Consciousness, self-consciousness, and introspective self-knowledge

By
Isabella Muzio

PhD in Philosophy
University College London
2005

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Acknowledgements

My thanks go to my supervisor Mike Martin as well as to Mark Sainsbury and Martin Davies for detailed comments and much encouragement on this and earlier versions of my work on self-knowledge. I would also like to give special thanks to Sophie Allen, Jerry Valberg, Barry Smith, Lucy O'Brien, Naomi Eilan, Heidi Maibom, David Harris, James Tartaglia, Mark Hammond, Guy Longworth, Scott Sturgeon, Matt Nudds, Matt Soteriou and countless others who have made inspiring suggestions and raised invaluable comments and criticisms at conferences, seminars, or just over coffee. Last but not least, my warmest thanks go to Stipo Androvic for proof reading the whole thesis, and to Lumi for just being her.

Abstract

We are, it seems, able to know a wide range of our own thoughts, beliefs, desires and emotions in a special immediate, authoritative way in which we are not able to know the mental states of others, nor indeed a certain range of our own such states. How is this possible? What *is* this special way we have of knowing a certain class of our own mental states? What, in fact, is the class of states of which we are able to have such knowledge, and, what is it about this class that enables us to know them in such a distinctive, authoritative way?

The broad aim of this thesis is to bring out, in answering these questions, an important point of intersection between issues about world-directed consciousness, self-consciousness and introspective self-knowledge.

More specifically, starting from the problem of authoritative self-knowledge, the aim of the thesis is threefold: to motivate, to articulate, and to expand upon a particular Sartrian solution to this problem, based on a view of our world-directed conscious states as being in some sense at the same time states of implicit or 'pre-reflective' self-consciousness.

In accordance with this threefold aim, the thesis divides into three parts as follows:

Part I begins with the problem of authoritative self-knowledge and the standard solutions on offer in the literature: inferential models, perceptual models, and constitutive accounts. It then suggests how a close examination of the shortcomings of these standard approaches ultimately points towards a solution along the above Sartrian lines, ie. based on an understanding of first-order consciousness as involving already itself an implicit form of self-consciousness.

Part II then focuses more narrowly on this notion of implicit self-consciousness, proceeding (a) to distinguish it first from other similar-sounding notions in the literature (ie. notions of 'non-conceptual' self-consciousness, higher-order-thought conceptions of consciousness, and constitutive accounts of self-knowledge), moving on then (b) to show how the notion introduced here, contra these others, can indeed provide the basis for a solution to the initial problem of introspective self-knowledge meeting all the desiderata on a successful such theory.

Finally, Part III takes on the more concrete issue of how such a form of implicit self-consciousness might, in practice, be seen to be involved in our two main categories of world-directed states, ie. in our cognitive states on the one hand (thoughts, beliefs, perceptual experiences), and in our emotions on the other (desires, fears, hopes, etc). This section of the thesis goes beyond mere concerns about the relation between an implicit form of self-consciousness and the problem of self-knowledge, drawing on both other parts of the philosophical literature and on various parts of the current psychological literature, to make not only more concrete sense of the view of world-directed consciousness here advocated, but to thereby show it to be also plausible independently from the theoretical considerations about self-knowledge initially driving it in this thesis.

Table of contents

Part I: Introspective self-knowledge

Introduction: The problem of introspective self-knowledge p.6

0.1 Immediacy, first-person authority, and immunity to non-cognitive error p.7
0.2 Conscious, non-conscious, and unconscious mental states – p.14

Chapter 1: Inference and inner sense. p.18

1.1 Self-knowledge by inference – p.19
1.2 Self-knowledge through 'inner perception' – p.22

Chapter 2: Non reason-based accounts. p.39

2.1 Artefact of grammar views – p.40
2.2 The weak constitutive view – p.55

Chapter 3: Our grounds for self-knowledge: an explanatory puzzle. . p.60

3.1 Justification and self-knowledge – p.60
3.2 Burge – p.65
3.3 Peacocke – p.72

Part II: Implicit self-consciousness

Chapter 4: Consciousness as self-consciousness. p.85

4.1 Perceptual accounts, constitutive accounts and the higher-order thought theory of consciousness – p.86
4.2 Pre-reflective self-consciousness, the threat of infinite regress, and the phenomenology of world-directed consciousness – p.89
4.3 Solving the explanatory puzzle – p.93

Part III: World-directed consciousness: cognition and emotion

Chapter 5: Cognitive states: self-consciousness, experience, and objectivity. p.100

5.1 Self-consciousness and experiencing the world as objective – p.101

5.2 Objectivity and self/world dualism – p.107
5.3 Circularity and infinite regress – p.116
5.4 'Pre-reflective' self-consciousness and 'non-conceptual' self-consciousness – p.119

Chapter 6: Knowing our own emotions. p.130

6.1 The puzzle – p.131
6.2 Knowing our own emotions and experiencing the world as objective–p.136
6.3 The cognitivist proposal – p.145
6.4 The way forward – p.151

Chapter 7: Expressions of emotion and desire. p.154

7.1 Outward expressions of emotion – p.155
7.2 Manifestations of emotion in consciousness – p.175

Chapter 8: Emotions, desires and hodological space. p.191

8.1 The appraisal theory of emotions – p.192
8.2 Solving the puzzle – p.200

Bibliography. p.212

Introduction: The problem of introspective self-knowledge

Unlike most other animals, we are not only able to have thoughts, beliefs, desires, and a wide range of emotions, but we are also able to *know* that we have them. Moreover, in the case of a wide range of these states (ie. our in some sense *conscious* ones), our knowledge of them is immediate, first-person authoritative, and immune to certain types of error, in a way that our knowledge of other people's states is not, and indeed in a way that our knowledge even of a wide range of our own states is not (eg. our unconscious or repressed beliefs, desires, fears, hopes, etc). As I sit here in the library for instance thinking to myself 'self-knowledge is non-observational in character' or visualizing with desire a cup of coffee I will be having later on, I am immediately able to know that I am reflecting on the nature of self-knowledge, and that I would like a cup of coffee, in a way that I am not able to know whether anyone else around me is engaged in a similar reflection or has a similar desire. Determining the latter would require my spending some time observing the behaviour of those around me, looking over to see what books they have laid out in front of them and ultimately, in the case of relatively complex states such as these, having to ask them. Similarly, gaining access to a wide range of my own mental states, that is, to my repressed or otherwise unconscious beliefs, desires and emotions may also require paying close attention to my behavioural patterns, if not years of psychoanalysis. Nevertheless, though clearly not all our knowledge of the contents of our own minds is in any way different from our knowledge of the minds of others, it still remains that in *some* cases we are able to know our own mental states in a way

that is distinctive, that is, in particular, in a way that is immediate, first-person authoritative, and immune to certain types of error. How is this possible?

To put things differently, the problem of introspective self-knowledge is this: we seem to have a way of knowing the contents of our own conscious minds that is unlike our way of knowing the minds of others, unlike our way of knowing our own unconscious minds, and indeed unlike any of the normal ways we have of acquiring knowledge whether of minds or of anything else, namely by inference and/or through our five senses. And yet, the possibility of this special kind of self-knowledge cannot be denied. I have no doubt for instance that I am now entertaining the thought of leaving this room to get a cup of coffee; it does not seem to me that I inferred this from any other beliefs of mine or from my behaviour (I was not looking at myself); and surely no one else here is better placed than myself to judge that I am entertaining this thought. Scepticism is not truly an option,[1] and the problem of self-knowledge thus not that of explaining *whether* but *how it is* that we are able to know a certain range of our own thoughts, beliefs, desires and emotions immediately, non-inferentially, authoritatively, and in a way that is immune to certain types of error.

In order to solve this problem, we will need to get clearer first about what exactly is at issue, in particular by considering the following questions: (1) In what sense and to what extent is our knowledge of our own minds immediate, authoritative, and not subject to error? (2) What is the exact class of states of which we are able to have such knowledge? and (3) What theoretical options are available to us for explaining this knowledge? Leaving aside the task of laying out the theoretical options available to us until the next chapter, the aim of this introduction will be to take on the preliminary task of articulating the problem itself, by considering questions (1) and (2) in turn. Thus, to begin with: what are the distinctive features of our knowledge of our own minds?

0.1 Immediacy, first-person authority, and immunity to non-cognitive error

[1] Even the most extreme of sceptics is going to have to provide *some* account of the distinctiveness of our knowledge of our own minds, perhaps not by explaining how this knowledge can be immediate and authoritative as a *kind* of knowledge, but at least why it might be so as a matter of *degree* by comparison to our knowledge of other things (eg. the minds of others, the outside world, etc.). This view, ie. Ryle's (1966), will be returned to in the next chapter.

0.1.1 *Immediacy.* Our knowledge of our own mental states is, first of all, *immediate*, at least in that coming to know introspectively[1] what we are currently thinking or what we believe or what we desire, or indeed in many cases how we feel about something, only seems to require that we ask ourselves the question. If I am asked for instance whether I am now thinking about my work or instead about what I am going to do later on, I only need to consider the matter in order to know which it is. That is, I do not seem to need to consult any evidence – at least not any evidence directly regarding my *mental states* (eg. behavioural evidence, psychological evidence, etc).

Having said that, I may in some cases need to consider evidence about how the *world* is, or at any rate I might need to *attend* to the world, in order to make a mental self-ascription. For instance, if asked whether I believe that it is raining, I may need to look out the window, that is, I may need to consider whether it is or is not *raining* in order to be able to say whether or not I *believe* that it is.[2] However, if no window is in the near vicinity, I am immediately able to say, not having any evidence either way (and hence not being prepared to say either that it is or that it is not raining), that I have no view on the matter. Similarly, if offered a slice of cake, I may first need to take a minute to contemplate the cake I am being offered in order to say whether or not I would like some. However, once I become drawn to (or repelled by) the cake upon looking at it, I am immediately able to say that I would (or would not) like a slice. In other words, once I have formed an attitude towards something, or once an object towards which I already had an attitude has come to my attention, no further step in most cases seems required for me to be able to say that I have the attitude in question towards it. Thus, although considering various facts or objects may be needed in order to come to know what mental states we are in, these will generally not be, on the face of it, facts about ourselves (ie. psychological or behavioural) but rather the facts or objects *towards which* we have particular attitudes. Admittedly, in many such cases, self-ascribing our own mental states will be far from immediate in the sense of 'instantaneous'. For example, if asked whether I like someone, I might need to go through a quite lengthy process of deliberation, of recounting and considering

[1] 'Introspectively' should at this stage just be taken as an intuitive label for 'in that special way we seem to have of knowing a certain class of our own mental states', whatever this *way* may in the end turn out to be (eg. a form of inference or perception), if indeed it turns out to be a *way of knowing* at all (cf. constitutive accounts of self-knowledge).

[2] See (Evans 1982, chapter 7, especially p.225)

various factors about this person's recent behaviour and personality. Even here though, the process of self-ascription would remain *unmediated* by any explicit consideration of myself or of my mental states.

In other words, in many cases of self-knowledge, ie. in the most puzzling cases (here referred to as the 'introspective' cases) our knowledge is, if not instantaneous, nonetheless somehow direct, in that we do not seem to need to consult any evidence explicitly about ourselves or our states of mind in order to know what states of mind we are in. Instead, we effect a direct move from contemplating the objects (loosely speaking)[1] towards which we have (or may have) an attitude, to the conviction that we do or do not have that attitude towards them. No intermediate step, that is, seems required between our attending to a particular aspect of the world towards which we have an attitude, and our being able to say whether we do or do not have that attitude.

Before moving on to the second distinctive feature of our knowledge of our own minds, it should be noted here that from this direct and non-inferential character of our introspective judgements it does not follow that these judgements are necessarily epistemically *ungrounded* or *baseless*. A judgement or a belief can in fact be both *non-inferred* from any other state yet *rationally based* on one. This may for instance be the case of our ordinary perceptual beliefs. My perceptual belief that there is a table in front of me is not inferred from my perceptual awareness of it, but may nonetheless be thought of as epistemically *based* on this state of awareness. My perceptual experience of the table may constitute the *evidence* on which my belief is based, and may thus in this sense be said to constitute a *reason* for my believing or judging that there is a table in front of me, although my coming to this belief on that basis involved no process of inference. Put differently, as I did not reason my way to this belief (ie. I did not come to it by way of any *argument*) it might make no sense to ask me to defend my belief that there is a table in front of me, yet this belief would still not be, from within my perspective on the world, rationally ungrounded. The mere fact therefore that our introspective judgements about our own thoughts, beliefs, desires, etc. are not the result of an inference from explicit premises regarding our mental states, and therefore may not admit of nor require any defence, cannot be taken in itself to count against reason-based approaches to self-knowledge in favour of *non*

[1] For immediate purposes, the objects of our attitudes can be thought of as either states of affairs, or as states of affairs as described in one way or another, or indeed as actual physical objects, imagined objects, and so on.

reason-based approaches. Claiming that our introspective self-ascriptions are rationally ungrounded (ie. not based on any evidence at all) is something that would require further argument.

0.1.2 *First-person authority.* Turning now to the second distinctive feature of our knowledge of our own minds, it is often noted that we also stand in a position of *authority* with respect to the contents of our own minds, that is, that we are somehow better placed (or at least significantly differently placed) than other people, to say what mental states we are in.

Yet, at the same time, our judgements about our own states of mind are neither infallible nor it seems incorrigible. Both minor mistakes and complete failures to know what we believe, desire, fear, etc. are possible, as testified for instance by common cases of self-deception. I may discover for example by suddenly catching myself behaving in a certain way, or by having this pointed out to me by a friend or an analyst (who might here be in a more authoritative position than I am) that I have an attitude of which I was completely unaware. More specifically, I may discover, say, through noticing my strong reaction at the mention of someone's name, that I have strong feelings towards this person, which I was either completely unaware of having (never having even considered the matter), or which I downright believed I *didn't* have, and hence had been self-deceived about. In yet other cases, I might misjudge my own attitudes as a result of some form of irrationality, or simply as a result of a failure to fully grasp the concepts I am using, or indeed on purpose, as in cases of lying. Our knowledge of our mental states is thus clearly neither infallible nor incorrigible, nor indeed always first-person authoritative. Mismatches between our first-order states and our second-order judgements are both possible and frequent. It still remains though that in *some* cases – those of interest to us here – we do seem to stand in a position of authority with respect to the contents of our own minds. This phenomenon will therefore still need to be explained. For example, if the thought is now occurring to me that space is not Euclidean, I seem to be far better placed to know that this thought is occurring to me than anyone else around me, no matter how attentive they might be to my behaviour. Or, suppose that I am consumed with anger at the person I am talking to, and, suppose additionally that I am quite good at controlling my outward behaviour. The person I am talking to might in such a situation be completely unable to tell that I am angry at them, while it would remain

entirely obvious to me. What might explain this? How should we understand this position of authority we seem to stand in with respect to certain contents of our own mind? In what sense, that is, or in virtue of what, might our knowledge of our own mental states be in some cases first-person authoritative?

What the fallibility of our mental self-ascriptions suggests is that the authoritative character of our introspective judgements cannot be a feature of these judgements considered merely as such, that is, considered just as particular kinds of judgements or statements – ie. as statements with either a certain *form* (the self-ascriptive form) or with a particular type of *subject matter* (ie. our mental states). Given that a self-ascriptive judgement with the same form and content (eg. 'I dislike Jones') can in some instances be authoritative (eg. when based directly on my contemplating Jones's character) and in other instances *not* be authoritative (eg. when reached inferentially on the basis of my having observed my own hostile behaviour towards Jones), the authoritative character of those of our self-ascriptive judgements that *are* authoritative, can clearly not arise out of their form and content alone, but must arise instead out of the particular *way* in which they, but not other self-ascriptive judgements with the same form and content, were reached.[1] Put differently, given the possibility and indeed existence of both authoritative and *non* authoritative judgements about our own minds, there can be nothing about self-ascriptive judgements *in general* that makes them authoritative. Rather, there must be something distinctive about the way in which some of them are arrived at – unlike, say, judgements of the form 'I am hereby thinking that p' which might count as authoritative simply in virtue of their self-verifying form. Our non cogito-like judgements about our own conscious states however, if authoritative, must it seems be so in virtue of being made on a certain special *basis*, or reached in a certain special *way* not available to others.

0.1.3. *Immunity to non-cognitive error.* Finally, the third often noted distinctive feature of our introspective judgements is that they are, though not infallible (even when authoritative), nonetheless immune to certain *kinds* of errors, in particular non-cognitive errors, or as Burge puts it 'brute' errors – ie. errors not due to

[1] For this kind of line of attack on the idea of the incorrigibility of a judgement considered *as such*, discussed though specifically in relation to perceptual judgements, see (Austin 1962, lecture 10)

any kind of cognitive deficiency (eg. irrationality, division of the mind, confusion, etc.) or conceptual deficiency (ie. the misapplication or incomplete grasp of a concept).[1] If I am consciously thinking to myself 'My keys are on the table' for example, I cannot it seems fail to know that I am thinking this. That is, if I consider the matter, I can first of all not fail to know that a thought is occurring to me, nor can I mistake this occurrent thought that my keys are on the table for, say, an occurrent thought that there is a book on the floor. In fact, if in these circumstances I were to assert 'My keys are on the table but I do not believe that they are', or if I were to believe that I am thinking about a book being on the floor when in fact I am thinking about my keys being on the table, there would seem to be reason to question either my rationality or my understanding of the terms I am using or my sincerity. There would *not* however seem to be any reason to question my rationality or conceptual competence if I were to say 'My keys are on the table but Jones does not believe that they are', or if I were to judge that Jones is thinking about a book being on the floor when she is in fact thinking about a set of keys being on the table (I may just not be very good at interpreting people). In other words, there appears to be a certain range of our own thoughts and other attitudes about which we cannot, without irrationality or misunderstanding, be mistaken about whereas we *can* be so mistaken about the attitudes of others, and indeed even about a certain range of our own attitudes (ie. our unconscious ones – in the Freudian sense).

Why this is so however may initially seem somewhat mysterious, as is perhaps best illustrated by Moore's paradox.[2] If we hear someone make a statement of the form 'p, but I do not believe that p' or 'I believe that p, but not-p', we tend to feel that the person in question is somehow contradicting themselves, even though, on the face of it, their utterance does not have the form of a straightforward contradiction – ie. 'p, but not-p'. Similarly, though perhaps even more puzzlingly, upon hearing someone say something like 'That chocolate cake is calling my name, but I do not want any' (ie. upon hearing them express a desire; yet deny that they have the desire just expressed), or upon hearing them self-ascribe a desire by saying 'I would love some of that cake' while at the same time looking at it with disgust and perhaps exclaiming 'yuk!' (directly expressing *lack* of desire for the cake), we would probably also tend

[1] See (Burge 1996)

[2] See (Moore 1942, pp.540-543 and 1944, p.204)

to feel that the speaker was either somehow contradicting themselves, or being irrational, or caught up in a conflict of desires – the desire for cake on the one hand and the desire not to eat it on the other. Formally however, again, no contradiction or conflict of any kind would be involved. How then are we to make sense of the strong intuition we nonetheless have that some kind of conflict or contradiction *is* involved in such cases? The sense we have that Moorean utterances (and by extension cases such as the above, involving the expression of a desire/emotion followed by the denial that one has the desire/emotion just expressed) involve some kind of contradiction, appears to be just another side of our earlier sense that our judgements about our own states of mind, whereas not our judgements about the states of mind of others, are immune to non-cognitive error. If someone misjudges what mental states they have, there must be, we feel, something cognitively wrong with them. A strong desideratum therefore on any satisfactory theory of introspective self-knowledge, or of second-order judgement, will be that it be able to account for this phenomenon – ie. either by solving Moore's paradox (ie. showing the above statements to amount to formal contradictions), or by explaining how/in virtue of what our introspective judgements might be immune to non-cognitive error.

Having said that, brute error does seem to be possible in some cases of mental self-ascription – eg. in cases of the self-ascription of unconscious states. Our judgements about our own unconscious states are generally arrived at inferentially from observations of our own behaviour, which seems to leave wide open the possibility of error *not* due to any cognitive deficiency, but, say, simply due to a failure to notice some crucial aspect of our own behaviour, or perhaps due to a failure to correctly interpret the behaviour observed. How should such cases be fitted into our picture of our introspective judgements as being immune to non-cognitive error?

There are, it seems, two possible ways of doing this. On the one hand, one could argue that the phenomenon of Freudian unconsciousness is a cognitive deficiency, and thus that although brute error *seems* to be possible when self-ascribing an unconscious state, this is only because the cognitive breakdown (ie. a division of the mind) ultimately underlying this possibility has not been taken into account. On the other hand, given that strictly speaking errors of misperception (or of misinterpretation – say, due to the application of an erroneous psychological theory) of one's own behaviour are not themselves cognitive errors, one might want to grant instead that some mental self-ascriptions are indeed *not* immune to non-cognitive

error, but point out that this still does not show our *introspectively-based* judgements not to be so immune. In fact the possibility of brute error when self-ascribing an unconscious state (ie. on the basis of inference from observation) only seems to confirm a point made earlier on: that the special status of some of our second-order judgements (ie. as immediate, first-person authoritative, and immune to non-cognitive error) should not be thought of as holding in virtue of some special feature of our second-order judgements considered *as such* (that is, as judgements with a particular form and/or content), but in virtue of the special *basis*, or special *way* in which they, but not other second-order judgements with the same form and content were reached. It is only because we are not able to come to judgements about our own unconscious states in any such special way that we have to resort, when self-ascribing these states, to ways of knowing (eg. inference from observation) that are *not* immune to non-cognitive error. Our 'introspectively' based judgements remain however so immune, and the problem of introspective self-knowledge is confirmed again to be that of identifying not some special feature of our second-order judgements, but some special *way* we have of knowing certain contents of our own minds. Before trying to identify what this special way of knowing might be however, one further preliminary question remains: what is the exact class of states of which we are able to have such knowledge?

0.2 Conscious, non-conscious, and unconscious mental states

The discussion so far has revealed that we are only able to know a restricted class of our own mental states in a way that is immediate, authoritative, and immune to non-cognitive error. Most obviously, this class appears to include our *occurrent conscious* states or *phenomenally* conscious states,[1] that is, our perceptual experiences (visual, auditory, etc.), occurrent thoughts (ie. acts of 'thinking to ourselves in words' or perhaps in images), conscious processes of reasoning, acts of assent, pangs of desire, episodes of emotion, episodes of visualization, etc. – in other words states that

[1] This term is introduced by Block in (Block 1995) where he distinguishes 'phenomenal consciousness' from 'access consciousness'. Though I will not be dividing mental states up quite in the same way as Block does, I will to some extent be borrowing his term 'phenomenal consciousness' in order to speak of mental states which, amongst other things, there is 'something it is like' for us to be in (see Nagel 1974).

are occupying our attention, or the objects of which are somehow present to our attentive minds. [1]

Less obviously, we also seem to be able to know immediately, authoritatively, and in a way that is immune to non-cognitive error a wide range of our *non*-occurrent, *not* phenomenally conscious, *dispositional* states, such as our ordinary beliefs (eg. the belief that my name is Isabella), desires (eg. to have a good life), and standing emotional states (eg. love for our family, dislike of a particular person, fear of spiders, etc.). Despite their differences with our phenomenally conscious states, these states can it seems still be thought of as conscious in the Freudian sense. They are, we might say, *non*-conscious but not *un*conscious, in that although not occurrent, they are nonetheless rationally integrated with the rest of our mental states, both conscious and non-conscious. They dispose us, that is, to behave in ways that make sense to us from our phenomenally conscious point of view, as well as from the point of view of our other non-conscious (but not *un*conscious) beliefs, desires and emotions. Moreover, although again not themselves conscious episodes, they can (and often do) come to manifest themselves in such episodes upon our merely turning our attention to their objects. My underlying fear of spiders for instance might come to manifest itself in my perceptual experience of a particular spider upon encountering it – eg. in this spider's coming to strike me as *to be avoided.*[2] Or, your underlying belief that London is the capital of the United Kingdom might come to manifest itself in this proposition's (ie. that London is the capital of the UK) coming to strike you as *true* upon your contemplating it in thought.

The states we are able to know in a special immediate, authoritative way thus appear to be unified in essentially two ways: (1) in their forming together a causally and rationally integrated point of view on the world; and (2) in being, if not themselves conscious episodes, states that can nonetheless come to manifest themselves in such episodes simply upon our turning our attention to their objects. Importantly, we have seen that in the case of our non-conscious dispositional states,

---

[1] I have left out mention of sensations, essentially on the grounds that the correctness of an account of our knowledge of our own sensations will vary greatly with what account one chooses to adopt of the nature of sensations themselves (ie. as mere phenomenal feels, as perceptual experiences of one's body, or indeed as some kind of world-directed states, etc.). Providing a full account of the nature of bodily sensations being beyond the scope of this thesis, the focus here will be only on our knowledge of our world-directed states, of which one may or may not want to think of sensations as constituting a part.

[2] How exactly we should understand the manifestation of emotions (or for that matter beliefs and desires) in first-order consciousness will be addressed in more detail as we proceed, and in particular in the third part of this thesis.

they *must* come to manifest themselves in such phenomenally conscious episodes (or we must at least turn our attention, be it in thought, perception, or imagination, to their objects) if we are to access them in any special, non-inferential, authoritative way. All the examples considered so far in fact seem to suggest that it is only via our attending to the objects of our non-conscious beliefs, desires, and emotions (and thus via some occurrent *episode* of perception, thought, imagination) that we are able to come to self-ascribe these states with authority. If asked whether you believe that London is the capital of the UK, or if asked whether you would like a slice of cake, or whether you are in love with someone, you will at the very least have to turn your attention to British geography, to the cake you are being offered, and to the person in question, in order to be able to say whether you do (or do not) have these mental states. On the face of it therefore, and for reasons that will of course need to be explained in what follows, the primary objects of our introspective awareness seem to be our occurrent conscious states, and only through having such states are we able to come to know our *non*-occurrent conscious states in a similarly special way.

In contrast to both of these types of states however (ie. conscious and non-conscious, or occurrent and non-occurrent) we also have states that are *un*conscious in the Freudian sense. These are states that are completely inaccessible to us except in the non-authoritative ways in which the states of others are accessible to us, and which moreover, in sharp contrast again to the states so far considered, are neither rationally integrated with any other of our mental states (whether conscious, non-conscious or indeed *un*conscious), nor capable of manifesting themselves in our occurrent experiences, thoughts, etc. upon our simply turning our attention to their objects. Suppose that I have a repressed feeling of resentment towards Mary for example (this feeling being repressed, say, because I feel guilty about feeling resentment). Upon turning my attention to Mary, I will in all likelihood not view her any differently (eg. as 'to be blamed', as 'to be hit', etc.) than I would if I had no unconscious negative feelings towards her. Or, suppose that I unconsciously believe that the butler committed the murder (this belief being repressed, say, because the butler is a good friend of mine, whom I would hate to think of as a murderer). Again, upon considering whether the butler did it, I would probably tend to feel that he did not, and may also be inclined not to view the available evidence as actually incriminating him. My desire not to think of my friend as a murderer might in fact have so transformed my view of the situation that I might find myself reading every

piece of evidence as, say, part of a plot to falsely incriminate the butler. Deep down however (ie. unconsciously) I may still believe that he did it, as would be manifest, if not in my conscious grasp of the situation, nonetheless in my outward behaviour. In order to come to know that I unconsciously believe that the butler did it therefore, or that I have a repressed grudge against Mary, I would have to pay attention primarily to any unusual patterns of behaviour that I might be exhibiting, or to any other potential outward manifestations of resentment (towards Mary) or of suspicion (about the butler) that might reveal to me that I have these states – eg. particular facial expressions, verbal slips, and so on. Turning my attention to the butler or to Mary or to any facts surrounding *them* alone would not be enough. Moreover, even when finally uncovered, such unconscious states would still stand apart from my conscious and non-conscious states, in that they would still remain somehow alien to me in the way that other people's states remain alien to us even when known – ie. in our still being unable to directly control or influence them through reasoning alone.

With these contrasts in mind between conscious/ non-conscious/ and unconscious mental states, together with a clearer understanding of the senses in which our introspectively based judgements can be said to be immediate, authoritative and not subject to error, we are now equipped to move on to ask and address the more substantial question with which we started and which is to guide much of the rest of this investigation: what *is* this special 'introspective' way we have of knowing certain contents of our own minds? Or, more precisely: how is immediate, authoritative, immune to non-cognitive error knowledge of our own conscious/non-conscious minds possible?

Chapter 1: Inference and inner sense

The recent literature on self-knowledge divides the theoretical options available to us as follows:[1] we must either know the contents of our own minds inferentially from observing our own behaviour, or we must do so directly observationally through some form of 'inner sense', or we must do so not in any *way* or on any special epistemic *basis*, but rather in virtue of the holding of some essentially *constitutive* link between our first-order conscious states and our second-order self-ascriptive judgements.

In the next two chapters, the aim will be to review these three standard theoretical options in their various forms, to consider their many virtues, yet ultimately to argue that none of them succeeds in the end in providing a fully satisfactory account of the kind of knowledge we *actually* have of our conscious states, that is, a kind of knowledge that exhibits all three of the distinctive features just discussed in the introductory chapter, and understood in the ways there specified. Having said that, a close examination of both the virtues of these three approaches and of the specific ways in which each of them fails, will prove invaluable in the path towards a more plausible account, not just by narrowing down the options (and revealing that the three standard ones are not in fact exhaustive), but also by bringing out more clearly the desiderata that any satisfactory theory of self-knowledge will have to fulfil, and in the end, we will see, by actually revealing to us the very kind of account that will be able to do so.

Taking therefore these three standard approaches in turn, the first two (ie. inferential and 'inner sense' models) will be examined in the present chapter, leaving

[1] See for instance (Boghossian 1989)

the complexities of third (constitutive accounts) to be examined in the next. So, to begin with: why might one want to think that our knowledge of our mental states is based on inference from observation of our behaviour?

1.1 Self-knowledge by inference

Clearly, we do sometimes come to know our own mental states on the basis of inference from observation of our behavioural patterns, much in the same way as (it is often assumed) we come to know the mental states of others. This is so in particular when coming to know our repressed or otherwise unconscious beliefs, desires and emotions. The important question for present purposes though is that of whether this might be how we come to know our own mental states in *all* cases, that is, even in those cases where no inferential process *seems* to be involved or where we do not *seem* to consult any behavioural evidence regarding our mental states in order to come to know that we have them.

According to Ryle,[1] inference from observation of our behaviour is how we come to know our own mental states even in the so called 'introspective' cases – eg. when self-ascribing an occurrent experience of a table, a simple thought about the weather, a desire for coffee, a feeling of irritation towards someone blocking our view at the cinema, and so on. If such an account is to be plausible though, it of course has to come with an explanation of why, if we do not *in fact* come to know our own mental states in any fundamentally different way from the way in which we come to know the mental states of others, it nonetheless *seems* to us that we do. Ryle explains this as follows. He acknowledges that there is a difference between our knowledge of our own mental states and our knowledge of the mental states of others, but, he suggests, it is a difference only in *degree* (ie. in how *well* we know our own mental states) not in *kind* (ie. in the *way* in which we know them). It is then this difference in degree, he suggests, that creates the illusion of a difference in kind. More specifically, the idea is that since we have observed our own behaviour for far longer and far more closely than we have anyone else's, we have become far quicker at recognizing patterns in what we do and say (creating the illusion of immediacy), far more accurate in our interpretation of these patterns (creating the illusion of immunity to error), and

[1] (Ryle 1966)

ultimately therefore on the whole more authoritative in our judgements about the contents of our own minds than in our judgements about the contents of the minds of others.

In other words, we are not on this view able to know our own mental states in any special *way* that is not available to others; but rather, given our greater proximity and familiarity with our own patterns of behaviour, we tend to be much faster, more skilled, and ultimately therefore more accurate in our deployment of the one way of knowing available to us (ie. inference from observation of our behavioural patterns) when applied to ourselves than when applied to others.

This account clearly has a number of advantages: it is simple, it is able to dispel the air of mystery surrounding the process of so called 'introspection', and, importantly, it is able to accommodate the *fallibility* of our knowledge of our own mental states even in so called 'introspective' cases. Despite these advantages though, it is ultimately unsatisfactory for a number of reasons.

For one thing, we are often able to know our own mental states even when no behavioural evidence is available to us for interpretation, let alone actually consulted. I am for instance clearly able to know that I am now thinking that there are nine planets in the solar system, or that I would like a cappuccino and a ham sandwich, although I am merely sitting at my desk, displaying no behaviour from which I could possibly infer that I have either of these mental states. In fact what kind of behaviour would that have to be? In most cases, the only behaviour that could possibly be fine-grained enough to allow us to know with precision what mental states we are in is verbal behaviour, yet clearly we are able to know much of what we believe, desire and feel, even when we are silent.

Secondly, this type of approach to self-knowledge leaves us without any clear explanation of why in some cases we are actually *not* authoritative in our judgements about the contents of our own minds although we have plenty of behavioural evidence available to us for interpretation. How is it, for instance, that although I may be behaving in a consistently hostile manner towards a colleague, I might not believe that I have any negative feelings towards them, while someone else, having observed me for far less than I have myself, would immediately be able to tell that I do not like this person? This cannot have to do with my not exhibiting sufficiently fine grained behaviour from which I could infer that I had this attitude, since by hypothesis, I am behaving in a manner consistently suggestive of my having the attitude in question.

In other words, in a number of cases, the authoritative or *non* authoritative character of our mental self-ascriptions appears to be entirely unrelated to the amount of behavioural evidence available to us for interpretation, let alone to our actual engagement or not in a process of inference from this behavioural evidence. Interestingly in fact, given the objections just raised, Ryle's suggestion that we engage in a process specifically of *inference* from observation of our behavioural patterns appears to be the *least* of his problems. The more fundamental problem with Ryle's view lies instead in his claim that observation of *externally available behavioural evidence* is required for self-knowledge – whether as the starting point of a process of inference, or as, say, a more *direct* ground for self-knowledge.

To make this latter point clearer, consider one of the background assumptions of Ryle's denial of a first-person/third-person asymmetry between our way of knowing our own minds and our way of knowing the minds of others: the assumption that, at the very least, our knowledge of *other* minds is based on inference from observation of their behaviour. But, is this so? Do we not (sometimes at least) come to know the mental states of others far more directly than that? One could it seems argue that although knowing the mental states of others requires observing their behaviour, it does not require engaging in any process of *inference* from these observations. Upon seeing someone smile for instance, we do not refer back to any beliefs we might have about the likely causal connection between smiling (physicalistically described) and being happy. Rather, we seem to perceive their smiling *directly* as suggestive or as expressive of their being happy, or perhaps even more strongly, directly as *meaning* that they *are* happy. People just *look* happy to us. Some forms of behaviour just seem to have come (be it through observed constant association, cultural conditioning, or as a result of innate hard-wiring) to immediately *evoke*, *mean*, or *designate directly* (ie. not by way of any additional assumption about their causal origin) certain types of mental states or indeed that someone *has* these mental states. The move from perceiving the behaviour of others to attributing mental states to them on that basis in other words may be a distinctively *non-inferential* one.[1]

Now, perhaps one could argue that our knowledge of our *own* mental states is also non-inferential in this way – ie. is arrived at not on the basis of *inference* from

[1] I am sympathetic to this view of our knowledge of other minds and will return to it in further detail later on in the thesis in the context of discussing our knowledge of our own/others' emotions. For more on this, see Chapter 7 below.

observing our own behaviour but *directly* on the basis observing our own behaviour, already mentalistically grasped. Although this may initially appear to be a more plausible view than Ryle's original inferential one, it does not seem to be sufficiently different from Ryle's proposal to escape the difficulties raised earlier on. In fact, in adopting this 'new and improved' version of a Rylean model, we would still be retaining the most problematic of Ryle's claims, namely the claim that we come to know our own mental states essentially on the basis of observing our own *behaviour* (even if perhaps more directly than Ryle thought) and that we therefore do not come to know our own mental states in any special *way* not available to others (though we may well be *better* at directly 'reading off' our own mental states from our own behaviour than the mental states of others from their behaviour). This, we have seen is problematic in that it leaves entirely unexplained (a) how, in numerous cases, we are able to know our own mental states with great accuracy despite clear lack of sufficient (if indeed *any*) behavioural evidence available to us – whether to be directly 'read' or as the starting point of a process of inference – and (b) why, in other cases, despite the abundance of behavioural evidence available to us, we are *not* authoritative in our judgements about our own minds (cases of 'denial' or self-deception).

Leaving aside therefore any account of self-knowledge according to which coming to know our own mental states requires, in one way or another, observing our own behaviour, let us turn to consider instead the more plausible suggestion that we might come to know our own mental states not on the basis of observing our own *behaviour* with the use of our *external* senses, but directly on the basis of observing our *mental states* through some form of *inner* sense or 'perceptual self-scanning mechanism'

1.2 Self-knowledge through 'inner perception'

The etymology of the term 'introspection' certainly suggests that, when 'introspecting' we are aware of our own mental states through some form of 'inner perception'. To 'introspect' is in some sense to 'look inside'. But, one might ask, in *what* sense? The verbs 'to perceive' and 'to see' are pre-theoretically used in a variety of ways, in particular to mean quite generally to 'know' or to 'understand', as when speaking of 'seeing' the truth of a mathematical proposition, or 'seeing' what someone means. Similarly, to 'look' may be used to mean 'to consider' or 'to think

about', as when saying that one is 'looking into some matter'. In its modern theoretical sense however, 'perception' refers fundamentally to *external sense perception.* Speaking therefore of 'introspection' in the theoretical context of philosophy can be seen to establish a (possibly false) analogy between so called 'inner perception' and external sense-perception, or between the knowledge we have of the contents of our own mind on the one hand and the knowledge we have of that which lies outside it, through sense perception, on the other. Asking whether this analogy is legitimate, is indeed the real philosophical question about whether or not introspective self-knowledge can be said to be perceptual, and will accordingly be our guiding question in this section.

The question before us is thus this: does our knowledge of our own mental states have the features distinctive of the kind of knowledge we have of the world, or even of ourselves, through sense-perception? In fact, what *are* the essential features of sense perception and of the knowledge we thereby gain? The discussion here will proceed in two stages. First, the question of whether the analogy between introspection and perception holds will be considered purely from the point of view of the first-person phenomenology of these two ways of knowing (introspection and external sense perception). From this perspective, it will be suggested that they do indeed have important features in common, giving the analogy between them an initial appearance of overwhelming intuitive plausibility. Then, the question of what more precisely is to be understood by 'perception' in this context will be turned to, that is, the question of what features of the kind of knowledge we gain through external sense-perception are essential to its qualifying as distinctively 'perceptual', and are therefore those to which introspective self-knowledge must conform if the analogy with perceptual knowledge is to have any content. On a number of these points, it will be argued that the analogy actually fails, some of which, it will be seen, are only damaging to some versions of the perceptual approach to self-knowledge, and others ultimately fatal to *any* perceptual account of self-knowledge.

Thus, in brief, the discussion here will be guided by the three following questions: (1) Why might one be initially drawn to thinking of introspective self-knowledge as perceptual? (2) What features must our knowledge of our own mental states in fact have if it is to properly count as perceptual? And finally, (3) *does* our knowledge of our own conscious mental states actually have these features? In the end, it will be argued that it does not, and that our knowledge of our own mental

states through so called 'introspection' is in certain very distinctive ways *non-*perceptual.

1.2.1 Starting from the purely phenomenological point of view, introspective self-knowledge appears to bear a number of striking similarities to perceptual knowledge.

To begin with, just like our knowledge of the world through sense perception, our knowledge of our own mental states is immediate and non-inferential.

Next, our judgements about our own conscious states appear to be, like our perceptual judgements about the world, somehow *rationally grounded* despite not resulting from inference. We do not, that is, seem to just 'find ourselves' in certain circumstances with a sudden impulse to self-ascribe a particular mental state. Rather, when self-ascribing a conscious attitude (or a consciously manifested attitude, in the case of dispositional states), our doing so appears to make rational sense to us from within our own first-person self-ascribing perspective.

Moreover, not only do we seem to make our judgements about our first-order conscious states on the basis of some sort of *rational grounds*, but we seem to do so specifically on the basis of some kind of *awareness* we have of ourselves as being in these states. For example, I might be thinking to myself 'it is a nice day' or 'I should get some work done', or I might be engaged in some process of reasoning about how to resolve some practical problem, etc., when suddenly, it occurs to me that I am thinking these thoughts or engaged in this process of reasoning (eg. I might just suddenly 'catch' myself in the act,[1] or actually turn to consider the question of what mental states I am in). Interestingly, when this happens, the information that we are thinking such and such does not usually strike us as a surprise. We tend to feel instead that we were somehow aware all along of ourselves having these thoughts or trying to resolve some practical problem; we were just not explicitly thinking about the fact that we were. Put differently, we do not generally feel that we just come up with our beliefs about what we are thinking, experiencing, feeling, etc. from nowhere. Rather, these beliefs seem to us to be based on our being (or having been) *aware* of ourselves as having the attitudes in question at the time of attending to their objects.

[1] See (Boghossian 1989, p.11)

Importantly, we not only seem to base our judgements about what we are *currently* thinking, feeling, etc., on some sort of awareness we have of ourselves as doing so, but we also seem to base our judgements about what we were *just a moment ago* (or much longer ago) thinking or doing or feeling, etc., on this same kind of awareness we then had of ourselves as being in these states. I may for example remember being engaged in a conversation with someone, or I may remember thinking that p, or being angry at X, although at the time I was not thinking about myself at all, but only about the topic of the conversation I was engaged in, or about the possibility that p might be the case, or about person X. In retrospect however, I may feel that I can immediately and authoritatively state what I was thinking, feeling, or doing at the time because I remember *thinking, feeling, doing so*, much in the way that I might later remember there being a green car parked outside my front door, although at the time I was not thinking about the car being there. I was however perceptually aware of the car, and, similarly it seems, at the time of conversing with my friend, or thinking that p, or feeling angry at someone, I was also (or so it intuitively seems) in some sense *aware* of myself as doing so/being so.

In sum, there appears to be on the face of it more to the idea that our knowledge of our own mental states is perceptual than a mere ambiguity in the verbs 'to look' and 'to see'. Having said that, appealing to a perceptual account of self-knowledge may not be the only possible way of explaining the above phenomenological features. It will in fact be suggested in chapter 4 that there is another way in which this might be done. It remains though that close attention to the phenomenology of mental self-ascription, ie. to our first-person subjective standpoint as self-knowers, is essential to fully understanding the phenomenon of introspection. Any account therefore that turns out not to be able to explain or otherwise accommodate the above considerations, ought, at best, to be regarded with some scepticism. In this respect therefore, perceptual approaches to self-knowledge certainly have at least one strong intuitive advantage. This however, is by no means sufficient for drawing the conclusion that introspective self-knowledge is, in fact, a form of perceptual knowledge. Other crucial criteria need to be met.

1.2.2 Amongst the may distinctive features of our knowledge of the world through sense perception, the following seem to be some of the most essential, and potentially most threatening to the analogy between perception and introspection.

(1) First, it is sometimes claimed that perceptual beliefs are beliefs only about the intrinsic, non-relational properties of that which they are about, or at least that this is so when that which they bear a relation to is not also being perceived or otherwise known. To borrow an example from Boghossian, we can tell, by merely looking at a coin, that it is of a certain size, has a certain shape, is made of a certain type of metal, has a certain colour, etc. but cannot for instance tell what monetary value it has.[1] This knowledge needs to be inferred from some other information about how the coin's intrinsic features relate to the possession of monetary value. Or, to borrow an example from Davidson, we cannot tell just by inspecting a burn on someone's skin whether or not it is a *sun*burn.[2]

Following on this idea, a recent objection against perceptual accounts of self-knowledge has it that we cannot possibly know the contents of our own minds perceptually, since, given a plausible externalism about mental content, the contents of our world-directed attitudes are not intrinsic, non-relational properties of these states.[3] In immediate response to this however one might point out that this objection assumes both a very narrow conception of what an attitude is, and a very narrow conception of what 'inner sense' might be. It assumes first of all that a mental attitude is an internal state of a person that happens to also have certain relational 'content' properties, rather than something that is *intrinsically* a relation to the world, and therefore the content properties of which could, conceivably, be perceived. Secondly, it assumes that 'inner sense' is necessarily to be thought of as 'inner' in the sense of being directed inward towards our *heads*, rather than inward towards our *perspective*, a perspective that may also be thought of as 'reaching out into the world'.[4] If looked at in this way, the relational nature of our attitudes need not it seems pose any imminent threat to the conception of self-knowledge as perceptual.[5] A different

---

[1] (Boghossian 1989, p.16)

[2] (Davidson 1986)

[3] See for instance (Shoemaker 1996) and (Boghossian 1989)

[4] Shoemaker considers, in a footnote (1996, p.212), the idea that 'instead of thinking of a belief as something internal to the person, and its contents as constituted by its relations to other things, one could think of it as 'reaching out into the world''. However, he then dismisses the thought that this might make inner sense models more plausible, because, it seems, he continues to assume that 'inner sense' is 'inner' in the sense of being directed inward towards our heads, that is, that 'inner sense' must be a way of seeing internal states of a person – he writes '...how could inner sense reach out into the environment?' Why, one might reply, could it not?

[5] The more serious threat thought to be posed by externalism is not one that applies restrictedly to perceptual accounts, but to the authoritativeness of self-knowledge in general – thereby motivating Rylian types of positions. I will not be going into this issue here, essentially because, in agreement with (Burge 1988; 1996), I take the issue

feature of perception, however, that does seem to make the analogy with introspection somewhat more problematic, is the following:

(2) In perceptual knowledge, there is, one might argue, always an intermediate perceptual state, or sense-impression (visual, auditory, proprioceptive, etc.) distinct from the perceptual belief or knowledge it gives rise to, as well as distinct from the object of perception that causes it.[1] My perceptual belief that there is a cup of coffee in front of me, for instance, is based on my being perceptually aware (through sight, and perhaps also smell) of there being a cup of coffee in front of me. My awareness of the coffee, is in turn somehow caused (let us assume) by the presence of the object. Similarly, my knowledge through proprioception that I am leaning forward rather than standing upright, is based on a distinctive sensation (a sense of a particular kind of imbalance; an impression of leaning forward), that is itself somehow caused by my body's being in a certain position. In other words, in both proprioception and more straightforward cases of external sense-perception, our perceptual beliefs seem to always be based on some intermediate informational state of awareness of that which they are about, distinct both from our belief and from their object. One might in fact argue that there *must* be such a distinction if such things as the possibility of misperception and the possibility of disbelief in one's senses are to be accommodated. Insofar as it is possible to disbelieve one's own senses, ie. to see something without believing it, one's beliefs and one's sense experiences cannot possibly be one and the same state; they must be distinct. Likewise, given the possibility of misperception (eg. the brute misidentification of an object, or the misperception of some of its properties), one's perceptual experiences and the objects thereby perceived must also be distinct. No mismatch between them would otherwise be possible.

But now, if this is an essential feature of perceptual knowledge, we have here an important point of dis-analogy between perceptual knowledge and introspective

---

of externalism/internalism about mental content to be somewhat orthogonal to the debate about a certain kind of self-knowledge. Externalism, it seems to me, may be relevant to questions about whether we are authoritative in our knowledge of what the meanings of our words *consist in*, but not so much to that of whether we are authoritative in our knowledge of what we are thinking, what we believe, how we feel, etc. *understood in our own terms*. In the former sense, we may well not be authoritative, and in the latter sense, although we are authoritative, the truth or falsity of externalism is to a great extent irrelevant to this being the case. To take an example, whether my term 'water' refers to water (H2O) or twater (XYZ) does not take away the fact that I am authoritative in my statements of the form 'I would like a drink of water', or 'I am thinking that water is wet'. Whatever my use of the term 'water' picks out in my first-order thoughts, etc., this same 'stuff' will be picked out by my use of the term in my second-order judgements.

[1] See (Shoemaker 1996, p.205)

self-knowledge. In self-knowledge, most would agree, and this despite some of the phenomenological features mentioned earlier, that there is no separate informational state (an experience of our first-order states) in between our first-order states and our second-order judgements about them, caused by the former and rationally grounding the latter.[1]

There are however two ways in which one might try to bypass this dis-analogy: (a) one might deny that there actually is a dis-analogy with introspection here, and argue that our introspective beliefs *are* based on a kind of experience of our conscious states, namely on the 'phenomenal feels' or 'what it is like' properties associated with having them, by which we might be able to 'sense' that we have them. Or, (b) one might deny that perception actually involves the existence of any perceptual states distinct from our perceptual beliefs.

To start with (a), this idea, if correct, would have the great virtue of providing us with an account of what it is about an attitude of ours being *conscious* (or consciously manifested in the case of dispositional states) as opposed *un*conscious, that enables us to know that we have it in a special way in which we are not able to know our unconscious attitudes (these not being accompanied by any phenomenal feel by which we could sense them). On closer examination though, this view turns out to be unsatisfactory for the simple reason that it is not clear at all how there could be a recognizable, and sufficiently fine-grained phenomenal feel associated with each attitude which could allow us to know that we had it. Consider my present thought that space is not Euclidean. This thought would have to always be associated with a specific kind of phenomenal feel which would be recognizably the same on each occasion in which this thought occurred to me, and which I would recognize even on the first occasion on which I thought it. It is not clear at all however that any purely phenomenal feel associated with any such complex thought – let alone with the conscious manifestation of a complex underlying desire or emotion (eg. guilt about having said X to Y) – could possibly be focused enough, or fine-grained enough, to ground a belief that I have exactly that attitude and no other.

Turning therefore to option (b), on some views of perception, in particular Armstrong's,[2] what is central to a belief's being perceptual is not that there be some

[1] See for instance (Shoemaker 1996, p.207) and (Burge 1996, p.105)

[2] (Armstrong 1968)

intermediate experience between this belief and that which is believed, but simply that there exist a reliable, contingent, causal mechanism linking the objects perceived to one's beliefs about them. On this view, perception amounts to nothing more than the acquiring of beliefs through a reliable causal mechanism. And, if so, holding a perceptual view of self-knowledge does not require being committed to the existence of separate phenomenal experiences of our conscious states on which our introspective beliefs could be rationally based. Armstrong himself in fact holds such a reliabilist view of self-knowledge (in addition to holding a reliabilist view of perception).[1]

The central claim of this model of self-knowledge is that there exists a reliable, but contingent, causal mechanism in our brain, a 'self-scanning process', that can be thought of as analogous to many of our other perceptual mechanisms (sight, hearing, etc.) and whereby our first-order conscious states directly cause us to have second-order beliefs about them. If one accepts this account of self-knowledge, combined with a purely reliabilist approach to perceptual knowledge, one can hold on to a view of self-knowledge as perceptual, despite its non-conformity to the above feature of having to occur via an intermediate informational state of 'awareness' of one's first-order states. Having said that, two questions remain: should we accept a purely reliabilist approach to perception? And, even if we do (or do not for that matter), what might be the advantages of thinking of our knowledge of our own minds on this model?

To start with the first question, a general problem with this whole approach to perception is that it does not seem to allow for the possibility of disbelief in one's senses. This is particularly problematic when thinking about ordinary cases of *object* perception, where disbelief in one's senses clearly does seem to be possible. I may in certain circumstances mistrust what I see (perhaps because I believe that I have been given some drug), and may therefore not form any beliefs corresponding to my perceptual experiences. Seeing, in this case, would not involve *believing*. Moreover, a purely reliabilist approach to perception seems unable to make room for the fact that, from the phenomenological point of view, when we make a perceptual judgement about something, we do not seem to just find ourselves with an impulse to make it, in the way that we just 'find ourselves' having perceptual experiences through no

[1] (Ibid)

rational choice of our own. Our perceptual judgements are judgements that, we generally feel, make rational sense to us. They are judgements that we feel we could *withhold* were we to have reason to mistrust our senses, in a way that we could not, out of any rational motive, withhold a perceptual experience if our eyes are open. A purely reliabilist account of perception thus seems unable to make room for the important phenomenological differences between seeing and believing, and is thus perhaps not the best way of thinking about perception.

Of course, even if this is granted, pure reliabilism might still be the right approach to *intro*spection. Upon further consideration though, it is also unclear what advantages there might be to thinking of our knowledge of our own minds on this model, even if it *could* be thought of as a genuinely perceptual model. As seen above, the main appeal of thinking of self-knowledge as perceptual was that it made it come out as *reason-based* (in the sense of making rational sense to us from our own point of view)[1] in accordance with the phenomenology of many common cases of introspective belief. But now, if to adopt a perceptual model of self-knowledge is to adopt a purely reliabilist, non reason-based account (in the relevant sense), we are left without any clear reason for preferring a perceptual model of self-knowledge to a non-perceptual one – other than, perhaps, the fact that perceptual accounts, whether reason-based or purely reliabilist, seem to be better able to accommodate the fallibility of self-knowledge than (we will see) certain non-perceptual accounts are. This however can only constitute a reason for preferring a perceptual account if it turns out that the kind of fallibility to which our knowledge of our own minds is subject can be properly likened to the fallibility of our senses. Let us therefore consider next the analogy between perception and introspection on the question of fallibility.

(3) In perception, the objects of perception and perceptual knowledge (eg. a table) bear no constitutive, conceptual or rational relation to anyone's awareness of them or perceptual beliefs about them (eg. one's experiences or beliefs about the table). That is, in perception, the relation between awareness and object of awareness

[1] One might hold a view of reasons according to which reasons are just to be reduced to probabilistic links between beliefs formed and facts the beliefs concern, in which case pure reliabilism could be thought of as a reason-based account. I am not sure however how this could possibly capture the intuitive idea behind our concept of a reason, and, more to the point, how it could capture the sense we have when making a judgement about the world, or when self-ascribing a mental state, that we do not just 'find ourselves' making this judgement, but that doing so somehow 'makes sense' to us from within our own point of view. For a more in depth discussion of what we should understand by a 'reason' or 'rational ground' in this context see the beginning of chapter 3.

is a purely causal relation between two ontologically distinct and conceptually independent existences. Brute errors of identification, due to no cognitive failure of `our own, are therefore always possible.

Now, according to perceptual accounts of self-knowledge, we have, in addition to our external sense perceptual mechanisms, an 'inner' (and also purely causal) perceptual mechanism in our brain by which we can directly perceive our own mental states, in an analogous way to the way in which, through our external senses, we can directly perceive outside objects. Brute error, on such a model, is thus also possible in self-knowledge. But, as seen in the introductory chapter, brute error of the kind possible in perception is *not* possible in introspective self-knowledge. It seems impossible that one could for instance mistake one conscious thought for another, or a belief that p for a desire that q, or say happiness about having finished one's exams with anger at person X, in the way that one might, due to no cognitive failure or perceptual malfunction, perceptually mistake one object for another (completely different one), say some clothes hanging on a chair in the dark for a seated person. Similarly, it seems impossible that one might (again due to no cognitive deficiency, confusion, etc.) fail to be able to say what, or whether anything at all, is currently going through one's own mind, in the way that one might fail altogether to be able to say what, or whether anything at all, is directly in front of one, due to some very thick fog for instance.

To sum up, given on the one hand that it seems to be of the very essence of perceptual knowledge that brute non-cognitive error be possible, yet precisely distinctive of introspective self-knowledge that it is *immune* to such error, self-knowledge cannot it seems be a form of perceptual knowledge. The perceptual approach to self-knowledge in other words collapses when it comes to accommodating the third distinctive feature of our knowledge of our own minds discussed in the introductory chapter, ie. its *immunity to non-cognitive error*.

A perceptual theorist of self-knowledge could of course try to reply that self-knowledge is only immune to non-cognitive error because our inner scanning mechanism is far more accurate and reliable than any other of our perceptual mechanisms, and because, moreover, there just happen to be no external factors (the equivalent of lighting conditions, etc. in the case of visual perception) in our heads, that might interfere with our perception of our mental states. The difficulty with this response though is that it is unable to explain why all errors in self-knowledge are due

to some cognitive failure or other: irrationality, division of the mind, etc. and never occur in the absence of such failures.

Consider the following examples of failures in self-knowledge. People with split personality disorders might fail, in one of their personalities, to be aware of a conscious thought had by them when in another personality, yet we would clearly not diagnose these failures as cases of benign misperception, but rather as cases of division of the mind, that is, as cognitively pathological cases. Or, consider cases of schizophrenics who complain of 'thought insertion', that is, who deny that they are responsible for the thoughts they are having, or who deny that it is *they* who are the authors of the thoughts they are introspectively aware of. Again, although we would seem to have here cases of downright failure to identify oneself as the *subject* of one's thoughts, we would not classify such failures as mere errors of identification made by perfectly rational subjects who, say, were not looking closely enough. Instead, we would attribute such errors to serious failures in their rational thought processes. More common examples of error in self-knowledge may also be found in ordinary cases of self-deception. Leaving in fact aside cases which we already know to be pathological, if someone were to come up to us and sincerely assert such things as 'It is 12 o'clock but I do not believe that it is', or 'Someone wants a cup of coffee but I do not know whether it is I who have this desire' or 'I am not sure whether it is now occurring to me that it is raining or whether I am wondering about the nature of self-knowledge', we would immediately assume that there was something wrong either with the rationality of their thinking or with their understanding of what they were saying. We would not take such errors to be cases of misperception without irrationality or conceptual deficiency.

In other words, in self-knowledge, we do not seem to have any cases which we would classify as simply failures of some inner scanning mechanism without irrationality or other cognitive breakdown, whereas in sense perception, almost all cases of failure in perceptual mechanism that lead to erroneous belief, we would *not* classify as involving any irrationality, conceptual incompetence, or other cognitive deficiency. If there is an inner scanner therefore, its proper operation is far more closely tied to the ascription of rationality than in the case of sense perception. This in fact suggests that there must be, in introspective self-knowledge, a *rational* relation between our conscious states and our self-ascriptive judgements about them, in addition to whatever underlying causal mechanism may or may not also be involved

at the sub-personal level. That is, having a first-order state must somehow constitute or give rise to a prima facie *reason* for believing that one has it, as only this could seemingly explain why a failure to believe that one has it should count as a *rational* failure.

(4) A fourth and final point of disanalogy between introspection and perception lies in the fact that speaking of perception or observation (understood as external *sense*-perception) generally implies the idea of a perspective or point of view *on* something, on something that lies *outside* the observing perspective. If so however, one might wonder whether it actually makes sense at all to think of our knowledge of the contents of our own minds, that is, of the contents of our *own knowing perspective*, on the model of a kind of knowledge that is, of its very essence, knowledge only of that which lies *outside* our knowing perspective. This worry is very much related to point (3) just made above.

How, one might wonder, could we perceive a mental state of ours *from the inside* so to speak, that is, from the very same point of view from which we have it? If perception involves having a point of view *on* something, looking into our own mind or into our own point of view, would it seems have to involve having a point of view on our own mind, thereby creating an immediate distance between our observing perspective and the perspective of the states observed. And, if we then tried to look into our *observing* perspective, another dissociation would occur between *this* perspective and the new observing perspective, and so on *ad infinitum.*[1] This is a problem that many of us may feel the pull of, yet it is somewhat obscure what we should make of it. What would it be in the first place to know one's own mental states as the subject of these states or 'from the same point of view' as these states or 'from the inside'? And, why could such knowledge not be perceptual? One way of getting at this idea is by first considering the notion of a 'mind' or a 'perspective'.

A 'mind' or a 'point of view' might, in this context, be taken to consist essentially in a coherent system of rationally related attitudes, that is, in a system of states that form together a single unified picture of (and stance towards) the world, and that not only fill the same logical space (in the way that two different people's attitudes may also do) but that are also causally and/or in some other way

[1] For a discussion of this phenomenon/intuition, although not specifically in relation to perceptual accounts of self-knowledge, see (Sartre 1969, chapter 2, section III)

explanatorily related. Particular beliefs, desires and emotions within the same system, and with appropriately related contents, will in other words immediately explain, affect, or give rise to actions and other attitudes with relevantly similar contents, within this system. Given this basic understanding of a mind or a perspective, knowing one's own attitudes from the inside, or having direct knowledge of one's *own knowing perspective*, might be understood as a case of having a belief about an attitude, where one's belief and the attitude it is about form part of the same point of view in the above sense, that is, form part of the same rational/explanatory system, and bear therefore (amongst other things) a direct internal rational relation to each other.[1]

But now, if this is what introspective self-knowledge amounts to, then it cannot it seems be *perceptual*, since, as seen earlier, the relation between a perceptual experience and its object must be a purely causal, *non*-rational relation, if brute non-cognitive error of the kind that clearly *is* possible in perception is indeed to be possible. To put things differently, if the objects of our perceptual experiences formed part of the same point of view as our perceptual experiences or perceptual beliefs about them, they would have to bear some kind of *reason-giving* relation to these perceptual beliefs (in addition to any causal relation), characteristic of what it would be for them to occupy the same point of view, or for the transition between them to occur from within one same point of view. If so however, a failure to perceive these objects, or to form a belief corresponding to them, would constitute a failure to take account of one's reasons, that is, a failure of rationality. From this it would follow that error without irrationality would not be possible, yet being subject to such error is, we have seen, of the very essence of perceptual knowledge. Knowledge had from and about the same cognitive perspective can therefore not be perceptual.

But, perhaps we never do have knowledge of ourselves from the inside in this very strict sense, and if so, what we call *self*-knowledge or *introspective* knowledge,

---

[1] Of course, not all pairs of states within such a system will be *directly* related in any way. My belief that there is a table in front of me for instance and my belief that there is a chair in front of me are not *directly* rationally or causally related, yet they can still both be understood as being held from within the same point of view, say, by virtue of the fact that they both arise/are rationally grounded in one same perceptual experience I am now having of a table and a chair. Having said that, the relation between our first-order states and our second-order self-ascriptive judgements must it seems be a more *direct* relation than that. In fact, whatever one's views on self-knowledge, it is reasonable to assume that our first-order states (ie. the *objects* of our second-order beliefs) are somehow involved in the causal history of our second-order beliefs about them. And, if this transition from first-order states to second-order beliefs about them is to be additionally understood as occurring (being made) from within one same point of view (in the above sense), it will also have to be understood as being an *internal rational* relation, ie. one that it 'makes sense' to us to make from within our own self-ascribing point of view – and not one that we just 'find ourselves' undergoing.

may well still turn out to be just a kind of perceptual knowledge. A perceptual theorist could argue that our knowledge of our own mental states is a kind of knowledge that is somehow *distanced* from the perspective of our first-order states, that is, a kind of knowledge where our mental states are objectified as if they were someone else's states. One could in fact argue following Armstrong that '...it is only an empirical fact that our direct awareness of mental states is confined to our own minds. We could conceive of a power of acquiring non-verbal non-inferential knowledge of current states of minds of others. This would be direct awareness, or perception, of the minds of others. Indeed, when people speak of 'telepathy' it often seems to be this they have in mind'.[1] In other words, the idea is that there is nothing inconceivable about the possibility of knowing the mental states of others through some contingent, non-rational, brutely causal telepathic perceptual mechanism, so why would there be a problem with supposing that our knowledge of our own states is perceptual in this way? In fact wouldn't this make more sense than to think otherwise? Why should we think that our knowledge of our own mental states is a special kind of knowledge of these states had *from the inside*, that is, had in virtue of some direct *rational* relation holding between these conscious states and our judgements about them?

Well, there does not seem to be any problem with the supposition that we might be able to know *some* of our own mental states perceptually through some form of extra-sensory perception, such as some of our unconscious states for instance. However, the crucial question here is whether the knowledge we *actually* have when we 'introspect', given its peculiar features (eg. the fact that only certain types of errors are possible and not others), and given the uses to which we put it (eg. in reflective reasoning) could be of this kind. That is, do we in actual fact have knowledge of some of our own mental states *from the inside*, or is all self-knowledge had from a different perspective from that which it is knowledge about, and so a kind of knowledge that we could conceivably also have of the minds of other people?

There are, it seems, at least three reasons for thinking that introspective self-knowledge is knowledge of our own attitudes had from the inside. First, if it were not, it would be difficult to explain why, in self-knowledge, error is so closely tied to irrationality. Secondly, if self-knowledge were not truly 'inner' in this way, it would be difficult to explain why when we non-inferentially self-ascribe a conscious

[1] (Armstrong 1968, p.325). See also Churchland's discussion of telepathic knowledge in (Churchland 1991, pp.610-611).

attitude, say, a belief that p or a feeling of hatred towards someone, we generally do so with a certain commitment to the view that indeed p (eg. when saying 'I believe it is time to go'), or with a certain hostility in our voice (eg. when saying 'I hate you'). In such cases, we do not seem to just be reporting a belief or emotion in the non-committed way in which someone other than ourselves might do so by saying 'She believes it is time to go' or 'She hates you'. Thirdly, and perhaps most importantly, actual examples of knowledge of our own mental states had from the same perspective from which we have them (ie. from the inside in the above sense) can be found in our practices of critical reasoning.[1] To illustrate this latter point, let us digress for a moment and consider what it is to reason critically.

Following Burge, critical reasoning is essentially reasoning where a thinker:

(1) recognises her attitudes, reasons, and reasoning *as* attitudes, reasons, and reasoning,

(2) reasonably evaluates these attitudes, reasons, and reasoning by reference to rational norms, and

(3) where these reasonable evaluations constitute immediate reasons, and immediately rationally result in, explicit confirmation, review, or supplementation of the attitudes, reasons and reasoning reasoned about.[2]

Now, if I start thinking about the various beliefs I hold about the nature of introspective self-knowledge (stage 1), and begin evaluating them and considering whether they are reasonable beliefs to hold, thereby reaching the conclusion that one of my beliefs is unreasonable (stage 2), this evaluation, that is, my conclusion that one of my beliefs is unreasonable, will constitute an immediate reason for my dropping this belief (stage 3), in a way that my coming to this conclusion about someone else's beliefs would not itself alone, constitute an immediate *prima facie* reason for them to drop *their* belief. In fact consider the following situation: I am listening to a philosopher expressing her views on self-knowledge, and I thereby start reasoning about her views as a result of which I come to the conclusion that one of her views is unreasonable. My merely coming to this conclusion is clearly not in itself enough to make her have an immediate reason to change her belief. In order for my evaluation to

[1] See especially (Burge 1996). See also Shoemaker's discussion of reflective reasoning in (Shoemaker 1988)

[2] As Peacocke points out (1996), we do not always reason fully reflectively in this manner, and other animals most likely never do. The only relevant point here however (not that Peacocke denies it) is that *we do* sometimes engage in this kind of reasoning, or are at least able to.

result in a reasonable change in her views, I would first have to convince her of the truth of my evaluation. My coming to think it alone would not immediately rationally result in her changing her mind, in the way that my coming to this conclusion would itself have an immediate effect on what views *I* hold. My reasoning about her views would therefore not constitute a process of genuine critical reasoning in the sense defined above.

In other words, the point to take from this is that genuine critical reasoning seems to involve there being a certain *rational integration* of our first and second-order attitudes. They must immediately rationally influence each other. That is, the first and second-order states involved in critical reasoning must form part of the same point of view, since the two levels (first and second-order) in such reasoning, if such reasoning is to be possible, must stand in immediate reason-giving relations to each other.[1] If this is right, then it looks as though we do sometimes have knowledge of our own mental states from the inside, that is, from the same perspective as the states thereby known. Moreover, given that the judgements we make when we introspect are the very judgements we use in critical reasoning, our introspective self-knowledge *must* be of this kind, and so cannot it seems be perceptual.

To conclude, we have seen in this section that the terms 'introspection' and 'inner sense',[2] in suggesting an analogy between our knowledge of our own minds and our knowledge of the world through sense-perception, capture certain important aspects of the phenomenology of introspective self-knowledge, yet are ultimately misleading for a number of reasons. First of all, any perceptual account of self-knowledge that is not committed to the existence of inner sense experiences, distinct from both our first-order states and our second-order judgements about them, ends up turning into an intuitively implausible, purely reliabilist account, which cannot even so clearly be thought of as perceptual. Secondly, given on the one hand the nature of perception as essentially involving a dissociation between the observing and the observed perspectives, and given, on the other hand, the notion of 'inner' knowledge as a way of knowing the contents of our *own* observing perspective, we end up having

[1] I will return to discuss in more detail the relevance of critical reasoning to our purposes in chapter 3, in the context of discussing Burge's views on self-knowledge.

[2] 'Inner sense' is not always used to refer to a form of inner perception in the sense discussed in this chapter. Kant for instance, in speaking of 'inner sense' in the first Critique does not seem to have anything like a perceptual model of self-knowledge in mind. Despite the misleading term 'sense', 'inner sense' in Kant just seems to mean 'self-consciousness', or that primitive unifying self-awareness that comes with all other conscious states including perceptual states; a 'sense' so to speak that accompanies all others but which is not itself one of them.

to choose between taking introspective self-knowledge to be truly 'inner', and taking it to be truly perceptual, that is, between taking it to be knowledge had *from the inside* or knowledge had *by looking*. It cannot be both. Since its immunity to non-cognitive error together with the actual existence of practices involving a rational integration of our first and second-order states show it to be 'inner' in this sense, it follows that it cannot be perceptual.

Having seen that our knowledge of our own minds, given its distinctive features, can neither be based on observation of our behaviour, nor on direct observation of our mental states, the next option to consider is whether it might be distinctive specifically in not being based on anything at all, that is, in *lacking* reasons.

Chapter 2: Non reason-based accounts

The view that self-knowledge is not reason-based is shared by a number of positions in the literature, ranging from strong constitutive accounts, according to which it is somehow ontologically constitutive of believing that one has a particular mental state that one actually does have that state, all the way, via conceptually constitutive views, to purely reliabilist views, according to which our first and second-order states are both ontologically and conceptually independent, although somehow causally linked at the sub-personal level. The very different positions within this range share however a common commitment to the view that our immediate introspective avowals are not based on reasons, that is, that they are not rationally grounded in any way: they are neither based on other beliefs, nor on observation (whether of our behaviour or directly of our mental states), nor even on our self-ascribed states themselves.[1] Rather, on each of these accounts, there is something else (if anything at all) in virtue of which our mental self-ascriptions have their distinctive features of immediacy, authority, and immunity to non-cognitive error.

Having already discussed (and dismissed) pure reliabilism in the context of talking about perceptual accounts of self-knowledge, the focus in this chapter will be instead on the various lines of *constitutive* non reason-based approaches to mental self-ascriptions, to be divided here essentially into two categories as follows: (1) *artefact of grammar* views or *strong constitutive views*, according to which it is in one way or other *ontologically* constitutive of having a second-order belief that one

[1] In the case of reliabilism, this is of course only true of *certain types* of reliabilist positions, in particular *not* of those which hold reliabilism across the board, and therefore according to which to be reason-based is just to be produced by a reliable purely causal mechanism.

actually does have the corresponding first-order attitude or vice-versa, and (2) *weak constitutive views*, according to which it is *conceptually* constitutive of having a first-order attitude that one will generally tend to form a correct second-order belief about it.

Ultimately, it will be argued in each case that the position offered is unsatisfactory, and that an *epistemological* (ie. reason-based) approach to mental self-ascriptions must be returned to, though neither an inferential nor a perceptual one. Through a close examination nonetheless of the non reason-based approach in its various forms, a more plausible *fourth* approach (ie. additional to the three standard ones of 'inference, inner sense, or nothing' so far considered[1]) will be seen to emerge, intermediate between the second (ie. perceptual accounts) and the third (ie. non reason-based views). This is an approach that *will* be seen to be capable of solving our initial problem of introspective self-knowledge – ie. in being able to adequately accommodate all three of the distinctive features of our knowledge of our own conscious minds discussed in the introductory chapter. In so doing though, this 'intermediate' reason-based approach will also be seen to give rise to a new and deeper explanatory puzzle, which it will be the task of much of the rest of the thesis to address.

First though, let us run through the various possible non reason-based attempts to solve the traditional problem of self-knowledge, and see how none of these attempts, as they stand, are ultimately able to do so.

2.1 Artefact of grammar views

The fundamental claim of these views is that the immediacy, authority and immunity to error of our knowledge of our own minds is just the artefact of a grammatical misconstrual, a misconstrual either of *expressions* of mental states as truth-evaluable *assertions* about them, or of a mere *language game* as reflecting a language-independent reality and special way of knowing it upon which this language game is consequential, or finally, the misconstrual of what are in fact self-verifying second-order judgements as reflecting a special way of knowing the first-order states they are about.

[1] The slogan 'inference, inner sense, or nothing' is taken from Boghossian in (Boghossian 1989).

To be more specific, according to the first kind of artefact of grammar view, the 'expressivist view' as it will be called, we sometimes, though not always, say things like 'I am in pain' or 'I believe that p' or 'I am angry', etc. simply as alternative ways of saying 'ouch' or 'p' or shouting (ie. directly expressing anger).[1] It is in *these* cases, according to this view, that our mental self-ascriptions are immediate, authoritative, and immune to certain kinds of error. When our mental self-ascriptions are used as actual *assertions* on the other hand, they do not, on this view, have any of the special features we generally associate with first-person avowals, but are just as indirect, and based on exactly the same kind of evidence, as our judgements about the mental states of others are. The problem of how it is that we can have immediate, authoritative, and immune to non-cognitive error knowledge of some of our own mental states is thus on this view just an illusion that arises from mistaking uses of 'I believe that p' or 'I love X' as *expressions* for uses of them as *assertions*.[2]

According to the second kind of artefact of grammar view alluded to above, there is actually nothing there to be explained about why our first-person avowals have the distinctive features they have, or nothing there to be said about that in virtue of which our avowals have these features; they just do. That is, it is just part of our practices with the words 'believe', 'desire' 'love', 'anger' etc. that a person's immediate claims about her own mental states are taken as correct and authoritative in all cases in which there are not strong overriding reasons for rejecting them.[3] That is, on this view, what someone believes, desires, intends, feels, etc. is not to be *inferred* from her avowals (as one might infer from someone's shouting that they are angry), but indeed in part to be *identified* by what this person (when sincere) claims to believe, desire, etc.

On the third type of view listed above, held in particular by Burge, although solely with respect to strict cogito-like judgements, our judgements of the form 'I am

---

[1] This approach is sometimes taken to be Wittgenstein's (Wittgenstein 1953), and is defended amongst others by Heal (1994). Wright (1998) however denies that this is actually Wittgenstein's view. Whether or not Wittgenstein held this position though is not important for the purposes of the present discussion. For present purposes, it only matters that this is one possible non reason-based view, and one which is not without certain advantages.

[2] There are many ways in which this approach might be made to look more plausible, such as by saying, following Heal (1994), that these expressions of mental states are not *only* expressions, but are at the same time to be taken as self-descriptions of oneself as satisfying certain behavioural criteria. See (Heal 1998, p.21). Whatever the details might be however of any particular account along these lines, what I am interested in here is only the particular strategy that such accounts appeal to in order to explain the distinctive features exhibited by our non-inferential utterances or thoughts of the form 'I believe that p', 'I am angry at Y', etc., and whether this strategy works.

[3] For a discussion of this position, which Wright calls the 'default view', see Wright (1998)

hereby thinking that p' are immediate, first-person authoritative and immune to error simply in virtue of their self-verifying form. I cannot indeed be thinking to myself 'I am hereby thinking that there are physical objects', without in fact thereby thinking that there are physical objects.[1]

In sum, on one kind of strong constitutive view (the 'expressivist view'), it is constitutive of someone asserting non-inferentially, say, that they believe that p, that they do actually believe that p, because asserting this is just another way of expressing their belief. On another such view (call it the 'default' view), it is a basic unanalysable fact about our practices with the words 'believe', 'desire', 'anger' etc. that if someone non-inferentially, sincerely, and with understanding asserts that they believe that p or that they desire X or are angry at Y, then they do, in virtue of that very fact, count as believing that p, desiring X, being angry at Y. That is, on this view, uttering or being disposed to utter, say, 'I believe that p' is constitutive of believing that p. Finally, on yet another strong constitutive view (call it the 'self-verifying' view), thinking a higher-order thought involves quite literally thinking the corresponding lower-order thought. These being the basic claims underlying the various types of strong constitutive accounts of self-knowledge, let us now consider what some of the advantages might be of adopting one or other of these positions.

(1) To begin with, strong constitutive views have the advantage of providing a straightforward account of why it is that our mental self-ascriptions (or at least some of them) exhibit the features of non-inferentiality, authoritativeness, a kind of transparency, etc. If to assert that one has a particular mental state is also in some sense to express that very mental state, or if to believe that one has a mental state constitutes in one at the same time the having of that state, then obviously no inference from first-order state to a self-ascription of it is needed, nor can any third-person judgement about the same state equal the authority of first-person self-ascriptions of it.

(2) Next, the strong constitutive approach fits well (in its various forms) with the fact that when we sincerely and non-inferentially say such things as 'I believe that this is the right thing to do' or 'I hate X', we do so with a certain commitment to the view that this *is* indeed the right thing to do, or with actual hostility, which we do not do when saying such things as 'Jones believes that this is the right thing to do' or 'She

---

[1] See (Burge 1988)

hates X'. On strong constitutive views, according to which asserting things like 'I believe that p' or 'I hate X' either constitute in one the mental states self-ascribed (the 'default' view), or are expressions of these mental states (the expressivist view), there is no difficulty in explaining this. This in fact links up with our previous discussion (in the section on perceptual accounts)[1] about our first and second order states being held from within the same point of view. Adopting a strong constitutive view of self-knowledge according to which our first and second-order attitudes are not truly *distinct* attitudes, provides us with a simple explanation of how they can both be held from within the same point of view: insofar as they are not truly *distinct* attitudes, they cannot *but* be held from within the same point of view.

(3) Concerning the expressivist proposal more specifically, this view has the added virtue of providing an appealing solution to one of Moore's paradoxes that other strategies might seem unable to deliver. It can provide an explanation of why one seems to contradict oneself (or appears irrational) when asserting things like 'I believe that p, but not p' although it is perfectly possible that one may believe that p, and yet for it not to be the case that p, and moreover for there to be nothing necessarily irrational or contradictory about someone *else*'s judging this to be the case.[2] If the expressivist proposal is right though in suggesting that judging 'I believe that p' is in some cases just an alternative way of asserting 'p', it becomes immediately clear why, in these cases, these Moorean utterances are contradictory: they amount to asserting 'p, but not p'.

(4) Finally, a closely related advantage of this kind of non reason-based view (ie. the expressivist view again), is that taking our immediate avowals of the form 'I believe that p' to be mere substitutes for assertions of the form 'p', fits well also with the datum pointed out by Evans, drawing on a remark by Wittgenstein, that when asked whether we believe that p, what we do is not look at ourselves and consider any evidence regarding our beliefs, but rather, we look out at the *world* and consider

---

[1] See chapter 1 section 1.2.2 point (4)

[2] See (Heal 1994). Moore's other paradox concerns statements of the form 'p, but I do not believe that p'. This paradox, Heal grants, could be dealt with by appealing to the consciousness of our self-ascribed thoughts (where a thought's being 'conscious' is taken to consist in, or just to somehow involve, one's being aware of oneself having it). In this way, the utterance 'p', expressing a conscious belief that p, can be expanded into 'I believe that p', thereby generating the contradiction 'I believe that p, but I do not believe that p' which is of the basic form 'p, but not p'. Using this strategy to explain the second paradox 'I believe that p, but not p' however does not work. It only generates 'I believe that p, but I believe that not p' which is not itself a contradictory statement, or necessarily an *irrational* statement to make, but only an acknowledgement of the fact that one has contradictory beliefs.

whether or not p.[1] That is, if I am asked whether I believe that it is raining, I will look not at myself but out the window, and consider whether it is or is not raining. And indeed, if, following the expressivist, to say 'I believe that it is raining' is essentially to say 'It is raining', nothing should seem more obvious than that the evidence appealed to in order to make this avowal should be evidence regarding the weather.

In brief, artefact of grammar approaches to avowals seem to have much to recommend themselves. Having listed their many virtues however, it is now time to re-examine these points with a more critical eye.

(1*) Concerning the first supposed advantage of these views, it should be noted that the mere fact that strong constitutive views are able to provide some kind of account of the distinctive marks of first-person avowals (of their immediacy, authoritative character, and immunity to certain types of error) is not enough to tip the balance in their favour. It only puts them, for the time being, on a par with all other approaches that are *also* able to do so.[2]

(2*) Next, concerning the fact that asserting non-inferentially that one believes that p, or that one hates X, tends to be done with a certain commitment to the view that indeed p, or with hostility towards X, this again does not *need* to be explained by reference to any constitutive principle. Insofar as the belief that p or feeling of hatred one is self-ascribing is a belief/emotion one actually *has*, then of course one will tend to display, in self-ascribing it, a certain commitment to the view that indeed p, or a degree of hostility, without this commitment or emotion having to be *constitutive* of one's self-ascriptive judgement.

(3*) Concerning now the third virtue listed above of the strong constitutive approach, and of the expressivist view in particular, namely that of its being able to provide an explanation of why Moorean utterances of the form 'I believe that p, but not p' seem to be contradictory (or seem to involve some degree of irrationality), we have here again only a negative advantage if it turns out that the contradictoriness or

---

[1] See (Evans 1982, p.225)

[2] In their own ways, inferential models and perceptual models also had ways of explaining how our mental self-ascriptions have (or appear to have) the above distinctive features. Their ways of doing so though turned out not to be entirely satisfactory in the end. The same may yet be discovered to be true of the explanations provided by various versions of the strong constitutive approach, not to mention that there may turn out to be yet other accounts that are also able to accommodate these features, and perhaps to do so better. We will come to one such account at the end of this chapter.

oddity of such Moorean utterances can also be generated *without* appealing to any strong constitutive link between second and first-order states. Heal claims that it cannot, and is indeed lead to embracing an expressivist account of avowals essentially as a result of her attempt to solve this Moorean paradox.[1] Two points however can be made here to suggest that this advantage might not be quite so decisive.

(i) First, there *is* it seems another way in which this Moorean paradox could be dealt with, indeed a way that Heal herself briefly mentions but does not pursue in her paper.[2] The idea is that, given the datum that the evidence we appeal to in order to self-ascribe some of our mental states is not evidence explicitly about these states but evidence about the *world*, then, insofar as we self-ascribe, say, a belief that p on the basis of evidence we have for p, it follows that to say 'I believe that p, but not p' is in effect to be asserting 'not p' (the second part of the utterance) in spite of the fact that we are in possession of evidence for p, and have therefore immediate reason to assert 'p'. To take an example, a certain contradiction or at least irrationality (ie. failure to take account of the evidence one possesses) would indeed be involved if one were to sincerely assert 'it is not raining' (ie. the second part of the Moorean utterance) while looking out the window and clearly seeing that it *is* raining (ie. the evidence one supposedly appealed to to assert 'I believe that it is raining'). Thus, insofar as things are as they seem (ie. that we turn to the *world* for evidence about what we *believe*), the apparent contradiction or irrationality involved in making Moorean utterances of the form 'I believe that p, but not p' could, potentially, be also explained by way of an *epistemological* account of self-knowledge, according to which the way the world appears to us can actually provide us with direct grounds for making truth evaluable judgements not just about the world, but also about our *beliefs* about the world. The virtue that the expressivist approach to mental self-ascriptions therefore has of being able to solve this Moorean paradox is not again necessarily *exclusive* to it.

(ii) Next, not only does the expressivist view appear not to be the only one able to offer a solution to the above Moorean paradox concerning cases of *belief* self-ascription, but it is not clear whether it is able offer any solution *at all* to many

[1] See (Heal 1994)

[2] (Ibid, p.19)

parallel such paradoxes that arise in cases involving the self-ascription of desires and emotions.[1]

Consider for example a case of someone who is, say, caught up in a heated argument, and, upon being asked to calm down, shouts back: 'I am perfectly calm!' – thus simultaneously self-ascribing an emotion and (amusingly) expressing clear *lack* of the emotion just self-ascribed. Now, just as upon hearing someone utter 'I believe that p, but not p' (ie. self-ascribe a belief, then express clear *lack* thereof), we would tend to feel here (would we not?) that a rational person, one who is 'thinking straight', or in full grip of their cognitive faculties (ie. not subject to some, albeit temporary, cognitive/rational or conceptual failure) should know better, and indeed *would not* make such a blatant error about their current emotional state, even though they could well it seems, without any irrationality or other cognitive failing, make such an error about someone *else's* emotional state in uttering 'She is perfectly calm' immediately followed by (or simultaneously with) the 'she' in question (and/or indeed oneself) shouting (ie. expressing clear *lack* of calm).[2]

So, how might the appearance of irrationality involved in someone's shouting 'I am perfectly calm!' be accounted for by way of an expressivist construal of the mental self-ascription here involved? Well, with great difficulty it seems. The problem is essentially this:

Being careful not to formulate Moore's paradox in a way that presupposes already a particular strategy for solving it, there are generally speaking two possible ways of approaching it: (1) by trying to show the speaker to be (despite surface grammar) somehow *contradicting* themselves in self-ascribing a certain mental state

---

[1] One might object here that the expressivist approach (as put forward by Heal for instance) was never *intended* to solve paradoxes concerning the self-ascription of any states other than belief. It should be remembered though that our main concern here is not with any particular philosopher's use of the expressivist approach to avowals but with the expressivist strategy *itself*, as offering a possible way of solving the problem of introspective self-knowledge before *us* in this thesis – a problem which, as it turns out, does arise both for the case of certain self-ascriptions of belief *and* for the case of many self-ascriptions of desire and emotion. This, recall, is the problem of explaining how it is that some of our mental self-ascriptions can be (1) direct (ie. unmediated by any explicit consideration of ourselves or of our metal states, but seemingly based, if on anything, on considerations about (or attention to) their *objects*), (2) authoritative, and (3) immune to non-cognitive error – an immunity particularly well illustrated by our reaction to Moorean type utterances, and by our similar reaction to parallel cases involving the self-ascription of desires and emotions.

[2] In uttering 'but not p', as in shouting, one is of course doing more than just expressing one's *lack* of the belief that p, or *lack* of calm. One is also making a *positive* claim (ie. that the contrary of p *is* the case), and expressing a *positive* emotion (ie. anger or agitation). What needs to be brought out though, if the appearance of *irrationality* or *contradiction* involved in doing such things is to be explained, is a sense in which in saying 'but not p' or in shouting, one is *going against* something that one said (or against something that was involved in one's coming to say) 'I believe that p' or 'I am perfectly calm'. It is thus what one is *denying* or displaying *lack* of in saying 'but not p' or in shouting, rather than what positive belief/emotion one might be expressing, that is most relevant here.

simultaneously with (or immediately prior to) expressing *lack* of the mental state just self-ascribed – ie. the 'artefact of grammar' strategy; or (2) by trying to show the utterer to be *irrational* (ie. to be failing to take account of the evidence they possess, given the evidence they do possess in such cases) or cognitively failing in some way, in *believing* that they have a particular mental state (as exhibited by their self-ascription of this state) when clearly they (consciously at least – in the Freudian sense) *do not* have the state they believe themselves to have (as exhibited by their direct expression of *lack* thereof) – ie. the epistemological strategy.[1]

Now, adopting an expressivist approach to avowals in order to account for the appearance of irrationality or contradiction involved in our example of someone shouting 'I am perfectly calm!' would it seems have to involve: (a) adopting the *first* strategy, ie. showing the utterer of 'I am perfectly calm!' to actually be (despite surface grammar) *contradicting* themselves (ie. expressing something of the form 'p, but not p') in self-ascribing the feeling of calm simultaneously with shouting, and (b) showing them to be doing so (ie. to be contradicting themselves) in virtue of the fact that in uttering 'I am perfectly calm' they are not making a truth evaluable judgement but merely *expressing* the feeling of calm. The problem however is that, first of all, insofar as expressing lack of calm through shouting does not (plausibly) constitute making a *judgement* (ie. stating that something is or is not the case), there is no way in which it could be shown that doing so (ie. shouting) amounts to *contradicting* (ie. literally 'stating the contrary of') what one stated in self-ascribing the feeling of calm. Next, if one adds to this the expressivist claim that in self-ascribing the feeling of calm by saying 'I am perfectly calm' one is not making a truth evaluable judgement either but merely *expressing* calm, the problem gets even worse. Insofar as directly *expressing calm* does not again (plausibly) reduce to *judging* something to be the case, there is no way in which the expressivist construal of this self-ascription of emotion could help generate a contradiction between one's self-ascribing the feeling of calm and one's shouting.[2]

---

[1] Moore's paradox, recall, was brought in in chapter 1 primarily as a way of illustrating, or as an alternative way of capturing, the sense we have that our 'introspectively-based' (ie. immediate, authoritative) mental self-ascriptions (such as those made in Moorean type cases) are *immune to non-cognitive error*. This left open the possibility that our uneasiness about Moorean type utterances might in the end be due *either* to some underlying formal contradiction being involved in making Moorean type utterances, *or* to the fact that the error made in such cases (given the grounds one has available when non-inferentially self-ascribing a mental state) must be a *rational* or (in some other way) *cognitive* error.

[2] It might be possible in some cases to let out an emotion *through* an act of judgement. This however would not make it the case that to be expressing this emotion just *is* to be saying that something is or is not the case. For a

In other words, the only way it seems in which the expressivist approach to avowals could be used to successfully solve Moorean type paradoxes involving the self-ascription of emotions and desires is by making a highly controversial cognitivist assumption about the nature of our direct expressions of emotion, and thereby about the nature of emotions themselves, ie. as being, or as being essentially reducible to, types of beliefs or judgements.[1] Anything short of making this assumption would leave us here (with regard to our example) with a mere case of combined expression of two contrary emotions (eg. calm and anger or agitation); not with a Moorean paradox.

Of course, one could try to argue that cases such as the above are *not* genuine Moorean type cases of the kind here under discussion (ie. cases where we feel, despite surface grammar, that the person is somehow contradicting themselves or being irrational in what they are saying), but just cases where we feel (and perhaps find surprising) that the speaker is subject to two contrary emotions simultaneously (love/hate, calm/anger, etc.). Upon brief reflection though, it is clear that this is not what is going on in our example. Compare in fact our case of someone shouting 'I am perfectly calm!' while caught up in a heated argument, with a clear case of someone expressing (and subject to) both calm and anger. A scene from one of Coppola's *Godfather* films springs to mind here, where a figure of authority, while remaining perfectly calm (as expressed by their soft and controlled tone of voice) utters: 'You have disappointed me' or 'This has made me very angry' (or some other statement along these lines). Upon hearing such words, uttered in this manner, we would no doubt be far from amused. We would not take the person in question to be contradicting themselves or to be making a mistake about their current state of mind (eg. believing themselves to be angry when in fact they are not), let alone to be subject to some blatant form of self-deception, irrationality or other amusing cognitive lapse. They are both calm *and* angry. Our example of the person spontaneously shouting 'I am perfectly calm!' on the other hand could not be more different. In that example the speaker is clearly *not* calm despite their self-ascription, and, we feel, they ought to have known better. It remains therefore a clear instance of the above

---

detailed discussion of the various ways in which (and vehicles through which) emotions are expressed see chapter 7 of this thesis.

[1] The cognitivist theory of emotions will be discussed further (and dismissed) in chapter 6 below.

Moorean type paradox and *not* a case of simultaneous contrary emotions. The above difficulty for the expressivist view thus remains.

(4*) Finally, turning to the last virtue listed above (ie. point (4) above), this was, recall, the virtue that the expressivist view had of fitting well with Evans's datum about where we turn to for evidence when coming to self-ascribe a belief. If to self-ascribe a belief about the world is just to express that belief, then of course looking out at the world will be the right way of going about making a mental self-ascription. Two parallel concerns to those just raised in point (3*) above however can it seems also be raised here, suggesting that this advantage might not be quite so decisive either – in being neither *exclusive* to the expressivist view, nor transferring well to cases involving the self-ascription of emotions and desires.

(i) First, concerning Evans's datum for the case of *belief* self-ascriptions (ie. the datum actually discussed by Evans), it is not clear that expressivism is again the *only* view supported by this datum. As mentioned a few paragraphs back, it might be possible to find an *epistemological* account of self-knowledge according to which turning to the world also comes out as being the right way to go about coming to determine what world-directed beliefs one has. One such epistemological account will in fact be considered further below, namely a possible intermediate position between a perceptual account and a strong constitutive non reason-based view, according to which our mental self-ascriptions are ontologically *distinct* from our first-order states, yet *rationally based* on these first-order (world-directed) states, and thus according to which *having* a world-directed state (or a conscious manifestation thereof – and so in any case attending to the world) itself constitutes an immediate reason for *believing* that one has it.

(ii) More puzzling though again is how the expressivist view might be able to accommodate *at all* the fact (noted already in the introductory chapter) that we also seem to turn to the world for evidence about what *desires* and *emotions* we have. If asked for instance whether I would like a slice of cake, or whether I miss person X, I will tend to turn my attention primarily to the cake in question (perceptually) or to person X (in memory, imagination) rather than to myself, be it to my behaviour, or to facts about my psychology. How might *this* datum be explained on a strong constitutive view according to which utterances of the form 'I desire a slice of cake' or 'I miss person X' are just *expressions* of desire and emotion rather than truth evaluable judgements about these mental states? Again, it seems, at great cost.

If we are to turn to the world for evidence when making a self-ascription of emotion essentially in virtue of the fact that in self-ascribing this emotion we are not actually making a truth evaluable judgement about it, but merely *expressing* it, it must be assumed (implausibly again) that in expressing an emotion in this way we are in fact making a judgement about the world, ie. stating that something is or is not the case, that is, doing something that can actually be supported by *evidence*. Alternatively, an expressivist might argue that, in turning to the world to self-ascribe an emotion, we are turning to it not so much for *evidence*, but just to *trigger* an emotion which, having triggered it, we then come to spontaneously express in uttering something of the form 'I am angry', 'I am afraid', etc. The cost however of adopting this alternative expressivist explanation is not negligible either – it involves going against the very datum in need of explanation (ie. that we seem to turn to the objects of our emotions/desires for *evidence* or *confirmation* when making a mental self-ascription), and indeed against an important feature of the phenomenology of 'introspectively-based' self-ascription mentioned earlier, namely that when non-inferentially self-ascribing a mental state (be it a belief or an emotion) we do not just 'find ourselves' doing so, but doing so is something that, we generally feel, *makes sense* to us from within our own self-ascribing and outward looking point of view.[1]

To recapitulate, the virtues of the artefact of grammar views of avowals discussed here appear to be (a) neither *exclusive* to them with respect to cases of authoritative *belief* self-ascriptions; nor (b) to transfer well to cases involving the self-ascription of emotions and desires. In order therefore to decide in favour of, or to definitively rule out, any such accounts we will have to look elsewhere. In particular, we will have to turn away from considerations strictly about what decisive *virtues* various 'artefact of grammar' theories might have, and turn instead to consider the fundamental tenets of these theories themselves, and see whether they can stand up to scrutiny. Let us take them in turn.

Starting with the *expressivist* view of self-knowledge, one simple consideration seems to count decisively against it. We are able to know non-inferentially, authoritatively, and in a way that is immune to non-cognitive error, a number of our thoughts, beliefs, desires and emotions even if we are silent, in a dark

[1] See our earlier discussion of perceptual accounts of self-knowledge and of the phenomenology of the process of introspective mental self-ascription – ch. 1 section 1.2.1, as well as the end of section 1.2.2.

room, and, say, unable to move. This, recall, was a consideration that counted also against Rylean accounts of self-knowledge according to which we come to know our own mental states in no different way from the way in which we come to know the mental states of others, namely from observing our own behaviour (verbal or otherwise). The expressivist proposal may of course be seen to be an improvement on Rylean type positions in that it is actually able to accommodate the distinctiveness of our mental self-ascriptions at least in *some* cases, namely in those cases in which our mental self-ascriptions are being used as mere alternative ways of expressing the mental states thereby self-ascribed. Having said that, if our knowledge of our own mental states is taken to only have the distinctive features of immediacy, first-person authority and immunity to certain types of error in cases where this so called 'self-knowledge' is not actually *knowledge*, but mere *expression* of mental states, the awareness we are able to have of mental states of ours which we are not actually *expressing* still remains entirely unexplained.

So much therefore for the expressivist proposal. But, more interestingly, what about this whole general idea that there is an ontologically constitutive link between making an attitudinal avowal and having the corresponding first-order attitude? The problem with this approach more generally, is that it does not seem to leave any room for the phenomenon of *error* or *self-deception* about what we believe, desire, feel, etc., that is, for the fact (mentioned already in the introductory chapter – section 0.1.2) that our first and second-order attitudes can, and often do, come apart. Despite the immunity to certain types of error of a wide range of our self-ascriptive beliefs, we *can* sometimes have second-order beliefs *without* having the corresponding first order attitudes.[1]

One might try to reply to this objection by arguing, following Bilgrami for instance, that cases of error and self-deception are not actually cases where we believe that we have an attitude which we do not in fact have, but rather, cases where we happen to have both the attitude self-ascribed *and* an attitude which is inconsistent with it. To say however that the self-deceived son in Bilgrami's example,[2] who behaves *consistently* and *only* contemptuously towards his father (the only evidence of his believing his father to be a fine person being his asserting that he does),

[1] This point is emphasised amongst others by Boghossian (1989) and Martin (1998)

[2] See (Bilgrami 1998, in particular p.218)

believes both that his father *is* a fine person and that his father is *not* a fine person, is just, it seems, either to deny the obvious (ie. the possibility of self-deception), or simply to reiterate in different terms that he believes that he believes that his father is a fine person, but in fact does not believe that his father is a fine person, and hence is self-deceived.

Another possible line of defence of the strong constitutive thesis might be to argue that it is only constitutive of our self-ascriptions of our *conscious* (or non-conscious) attitudes that we have the attitudes self-ascribed, thereby leaving room for the possibility of being mistaken about a wide range of other beliefs, desires and emotions of ours (ie. our *un*conscious ones). The problem with this approach though, is that it seems to force us to say that it is ontologically constitutive only of *some* second-order beliefs of ours that we have the first-order states they are about, whereas it is not ontologically constitutive of other such beliefs of ours with the same content.[1] That is, for example, we end up having to say that in some instances, it is constitutive of my believing that I believe that it is raining that I actually do believe that it is raining, and that in other instances it is not. But now given that there seems to be no difference in the nature of the *second*-order beliefs in these two kinds of cases (other than the ad hoc difference that some are infallibly correct in virtue of some constitutive principle which does not happen to apply to the others), and given that the whole difference was supposed to hang on the (conscious or unconscious) nature of the first-order states self-ascribed, a more sensible approach at this point would be to say that our first and second-order sates are actually *distinct* states, and to try to see what it might be about a first-order states' being *conscious* (as opposed to *un*conscious) that connects it particularly intimately to second-order beliefs about it. This however is almost by definition not a line that defenders of a strong constitutive theory of avowals would want to take. Not doing so however, at this point, just seems to be to insist without argument that the constitutive thesis is right (perhaps because no better account seems to be at hand)[2], and in effect just to say that our second-order judgements about our mental states are infallibly correct in all cases (ie. those in

[1] This point is emphasised by Martin in (Martin 1998)

[2] In fact, if it turns out that none of the more explanatory theoretical options work, we may have to just admit defeat and resort to a quietist position of just describing how things are, and admitting that there is nothing more illuminating to be said. As Wright points out however, his view, which he calls the 'default view' is indeed a *default* view, that is, a view we can only be satisfied with if it is shown that no better or more illuminating account of avowals can be given. See (Wright 1998, especially pp.44-45)

which our second-order states constitute in us at the same time the states self-ascribed) except in those in which they are *not* infallibly correct (ie. those in which having the second-order beliefs does *not* constitute in us the having of the first-order attitudes self-ascribed).

This whole problem in the end links back, as anticipated in the introductory chapter, to an issue raised there about the exact sense in which our self-ascriptive judgements can be said to be immune to error (or for that matter immediate and first-person authoritative). It was argued, recall, that in light of the fact that some of our second-order judgements do *not* exhibit the special features of immediacy, first-person authority and immunity to non-cognitive error, these features cannot be taken to be features of our second-order judgements considered merely *as such* (ie. as judgements with a particular form and content), but must instead be thought of as features of some special *way* we have of knowing some, but not all, of our mental states. It is, it seems, a fundamental mistake of artefact of grammar approaches in general, to take the problem of introspective self-knowledge to be a problem about certain kinds of *statements*, namely 'avowals', or about knowledge with a certain type of *subject matter*, ie. our own minds.[1] The basic idea behind these views is that judgements about our own mental states are problematic because we take them to be authoritative, incorrigible, we do not ask people to defend them, etc. However, it is unclear that statements about our mental states taken *as such* actually *are* problematic, since in many cases (eg. when these statements are, say, about unconscious attitudes which we have come to know inferentially on the basis of behavioural evidence) there is no difficulty in explaining how they are possible. It is only in certain *specific cases* that our avowals are problematic, in that in these cases they appear to be non-inferred, authoritative, and it does not seem to make sense to ask us to defend them. The distinctiveness of these judgements therefore in certain cases, given that in other cases these same judgements (ie. judgements of the exact same form and with the same content) are *not* in any way distinctive, must therefore be a matter of *how*, in some cases and not in others, we are able to make them non-inferentially, authoritatively, and in a way that is immune to non-cognitive error. And, insofar as the question of *competence* arises, the possibility of authoritative self-knowledge raises traditional

[1] See in particular the way in which the problem of self-knowledge is set up by Wright (1998), as a problem about 'avowals', rather than as a problem about the way in which some of them are arrived at.

epistemological issues which cannot be dealt with through a purely metaphysical account.

Nonetheless, adopting a strong constitutive approach to attitudinal avowals might be appropriate when dealing with the narrower category of strict cogito-like judgements. Insofar as strict cogito-like judgements of the form 'I am hereby thinking that p' are self-verifying, their being non-inferred, first-person authoritative, etc. can be taken to be a feature of these judgements considered *as such*, rather than as reached in a certain way. However, following Boghossian,[1] it is clear that such an account is not sufficiently general. It takes us nowhere nearer understanding the immediacy, authority, and immunity to non-cognitive error of our knowledge of other attitudes of ours which are either not strictly contained in our self-ascriptions of them, or not strictly simultaneous with these self-ascriptions. If I was just a moment ago thinking to myself 'It is cold in here' I am immediately and authoritatively able to say that just a moment ago I was thinking this. Or, when I non-inferentially judge 'I would like a cup of coffee', my desire for a cup of coffee is in no way contained in my self-ascriptive judgement. I could be lying, or indeed I could be wrong. It may in fact turn out that once I get the coffee I realise that what I actually wanted was a beer.

Adopting a strong constitutive approach to avowals might thus be the right approach to *some* cases of self-knowledge, namely strict cogito-like cases, just as adopting an *inferential* account of self-knowledge may well be the right approach to other cases, eg. cases of knowledge of our own unconscious states, and indeed just as adopting a *perceptual* model of self-knowledge might be the right way of dealing with yet other cases, namely cases, were they to exist, of 'telepathic' knowledge of some of our own mental states. The problem however is that none of these approaches seems able to deal with the most common and most problematic cases of self-knowledge, namely those where we are able to know what we believe, desire, feel, etc. authoritatively, without having to infer this from explicit data regarding our mental states, and in a way that is, although not infallible, nonetheless not subject to certain kinds of error. Our initial problem of introspective self-knowledge therefore still remains, as does however one further possible standard line of approach to it: the 'weak constitutive' non reason-based approach.

---

[1] (Boghossian 1989)

2.2 The weak constitutive view

On this view, sometimes also called the 'weak special access functionalist theory',[1] it is conceptually (though not ontologically) constitutive of having a first-order state that one will generally tend to form a second-order belief about it when one considers the matter (ie. of whether one has it). In other words, the idea is that, following a functionalist theory of mind, according to which mental states are to be individuated by way of their functional role, that is, by reference to their relations to other mental states and to behaviour, self-ascriptive higher-order beliefs are amongst the mental states a lower-order state's relations to which are constitutive of it. For a state to be a *belief*, for instance, it must first of all tend to give rise to other beliefs, and tend to combine with desires and other attitudes with appropriately related contents to give rise to yet other attitudes and actions. The claim about self-knowledge then is that such first-order beliefs or indeed other first-order attitudes, not only have conceptually constitutive dispositional links to other first-order attitudes and to behaviour, but also to second-order beliefs about themselves. That is, for one to qualify as having, say, a certain first-order belief or simple emotion or desire, one must not only be disposed to form other appropriately related first-order beliefs and other attitudes, but also, in certain circumstances, to non-inferentially *self-ascribe* this belief/emotion/desire itself, that is, to form a *second-order* belief about it.

At first, this idea might seem to make a lot of sense, given that we would not generally be prepared to ascribe to someone, without thorough consideration of any overriding evidence, a mental state which they themselves did not believe themselves to have. In addition, this kind of constitutive view seems to have the advantage over strong constitutive views of leaving plenty of room for the possibility of error and self-deception, given its commitment to the ontological distinctness of first and second-order states, combined with the thought that it is constitutive of first-order states only that they tend *normally* (in circumstances of full rationality, reflectiveness, etc.) to give rise to second-order beliefs about themselves. There is therefore it seems something very appealing about this approach to introspective self-knowledge. It seems able to accommodate all three of the distinctive features of our knowledge of

[1] See (Fricker 1998)

our own minds without being subject to the problems faced by the other available standard accounts. Nevertheless, two features of it seem worthy of notice:

First, the supposedly constitutive feature of mental states that they tend to give rise to second-order beliefs about themselves needs to be relativised it seems to *conscious* mental states,[1] since many psychological states of ours *never* give rise to non-inferential self-ascriptions of them (eg. our unconscious states). It can therefore surely not be conceptually constitutive of *these* states, namely the unconscious ones, that they be dispositionally linked to second-order beliefs about themselves, although we would certainly still want to count them as genuine beliefs, emotions, desires, etc. In other words, the weak constitutive functionalist thesis can it seems only apply to *conscious* states, that is, it can only be conceptually constitutive of having a *conscious* state that one will tend to form a second-order belief about it. The basic idea behind this view therefore has to be that it is not conceptually constitutive of having mental states that they be dispositionally linked to second-order beliefs simply in virtue of their being *beliefs* or *desires*, but essentially in virtue of their being *conscious* beliefs, desires, etc. But now this might tempt one to ask what it is about a state's being conscious (as opposed to *un*conscious) that makes it the case that if one has it, one will tend not only to form other first-order states with appropriately related contents, but also to self-ascribe it. The importance of this question will become clearer as we proceed.

A second, and not unrelated noteworthy point is that when we consider the nature of the first-order states to which a certain other state's relations are conceptually constitutive of it, it is striking that these are all states that have relevantly similar contents to the one being individuated. That is, on most functionalist theories, the conceptually constitutive functional role of, say, a belief that p, is its tending to give rise to yet other beliefs and actions with relevantly similar or appropriately related contents. It is, for instance, part of the functional role constitutive of my believing that there is ice-cream in the refrigerator, that this belief will tend to combine with my desire for ice-cream, to give rise to an action of going to the refrigerator and getting the ice-cream. In other words, a functionalist individuation of mental states, on a theory according to which it is *conceptually* constitutive of mental

[1] Unless otherwise specified, by a 'conscious' state I will mean from here onward an 'occurrent or non-occurrent' conscious state as distinct from an '*un*conscious' state until we reach Part III of the thesis where more subtle distinctions will be relevant (the occurrent/non-occurrent distinction).

states *qua* mental (and not, say, *qua* physical) that they be related to other mental states and to behaviour, seems to be in effect an individuation of them by reference to what states are good reasons for having them, and what states they themselves are good reasons for having.[1] Thus, the idea of mental states having *constitutive* dispositional links to other mental states and to behaviour is the idea that mental states are to be individuated essentially in terms of what are typically good *reasons* for having them, or what attitudes they themselves are typically good reasons for having, and, these reason-giving relations can be made sense of, it seems, by reference to certain appropriate relations holding between their contents – eg. my fear of spiders will combine with certain beliefs about how to get away from spiders, to give rise to effective courses of action taking me away from spiders. It will *not* combine in this way with beliefs about, say, what there is in the refrigerator, or give rise to actions with regard to ice-cream.

The problem however now with the weak constitutive functionalist view of self-knowledge is the following. If the above is what it is for it to be conceptually constitutive of a mental state to be dispositionally related to other mental states and to behaviour, the next question is: given the unrelatedness of first and second-order contents,[2] how are we to make sense of the thesis that it is conceptually constitutive of having a first-order attitude that it will tend to give rise to a second-order belief about itself? Or, how are we to make sense of the idea, lying at the very basis of the weak constitutive theory of self-knowledge, that this constitutive link is on a par with the constitutive link our first-order states bear to other first-order states and to behaviour? There are, it seems, two ways in which one might go from here:

(1) One could try to explain how a first-order conscious state can stand in the same kind of relation to a second-order state as it does to other first-order states in relation to which it is conceptually individuated (ie. in some kind of internal *rational* relation), and this could be done perhaps by explaining what it is about a state's being

---

[1] The notion of a reason-giving transition will be discussed more fully in the next chapter. For present purposes though the idea is, briefly, that the notion of a reason-giving transition between two states can be understood roughly by reference either to logical relations (particularly when dealing with reason-giving relations between beliefs), or to some appropriate overlap, holding between their contents. For example, my believing that there is ice-cream in the refrigerator combined with my desire for ice-cream will *not* constitute an immediate reason for my starting to read a book on self-knowledge, although it *will* constitute a reason for my going to the refrigerator and getting the ice-cream. Similarly, my believing that p, combined with my believing that q, will *not* constitute an immediate reason for my believing that z, although it may well constitute an immediate reason for my moving on to believe that (p and q).

[2] First-order states are about the *world*, while second-order states are about *mental states*.

conscious that allows it to stand in such a relation to a second-order belief about itself. That is, one could try to explain how a first-order conscious state can stand in a *rational* or internal *reason-giving* relation to a second-order state despite the unrelatedness of first and second-order contents. To do this however would be to drift away from a non reason-based account.

(2) One could hold onto a non reason-based view, and argue that although the conceptually constitutive connection a first-order state bears to other first-order states happens to be associated with *reason-giving* relations holding between them, this just simply happens not to be the case of the constitutive connection between a first-order state and a self-ascription of it. If so, however, that is, if our first and second-order states are both ontologically and rationally unrelated, then it is not clear how the relation between them can be *conceptually* constitutive of having them. That is, at the very least, it cannot be constitutive of a first-order state *qua mental state* that it will tend to give rise to a second-order belief about it, though it may perhaps be constitutive of its physical realization in the brain.

If *this* however (ie. the latter) is what the view amounts to, it ceases to be clear how it is any better than an entirely *non-constitutive* purely reliabilist view of self-knowledge, given that it will end up being just as incapable as non-constitutive theories are of explaining, amongst other things, why we intuitively feel that it is a matter of *conceptual* necessity that one cannot have a conscious state without being disposed to self-ascribe it, or of explaining why Moorean utterances of the form 'p, but I do not believe that p' are *conceptually* odd, or indeed *contradictory*. In other words, if this (ie. (2)) is what the weak constitutive view amounts to, it is no more plausible than a purely reliabilist view of self-knowledge – which we have already rejected. If on the other hand it amounts to more than this (ie. (1)), then it is no longer a non reason-based view.

We have now reached the conclusion that introspective self-knowledge can first of all not be based on any of the normal ways we have of gaining knowledge, namely by inference from observation (or direct observation) of our behaviour or through direct 'inner observation' of our mental states. And, in this chapter, we have seen that the issue of the distinctiveness of our introspective second-order judgements cannot be a purely metaphysical issue regarding a certain class of judgements, as suggested by strong constitutive non reason-based accounts. A certain version of the weak constitutive approach however appears to be somewhat more promising, in that

it seems less subject to the problems faced by its stronger counterparts. However, this turns out to be so essentially in virtue of its not actually being a *non reason-based* view after all, but a view implicit in which is the idea that our mental self-ascriptions are based on reasons, not however on inference or observation, but directly on the mental states thereby self-ascribed. In other words, in spite of the failure of the three standard options of 'inference, inner sense or nothing', we seem to have here a possible intermediate position between the second and the third. This position however seems to immediately give rise to a further problem, which any satisfactory account along these lines is going to have to resolve, namely that of explaining how having a first-order conscious state can itself constitute a reason for self-ascribing it.

To put things differently, we are now faced with the following question, which is to guide much of the rest of this thesis: how can having an attitude towards the *world* constitute an immediate reason for believing something about our *mental states*, given that nothing about what mental states we are in follows directly from how things are in the world?

Chapter 3: Our grounds for self-knowledge: an explanatory puzzle

The central task of this thesis has shifted. Our analysis of the non reason-based approach in the last chapter, and through it the discovery of the only type of account seemingly able to offer a satisfactory solution to our initial problem of authoritative introspective self-knowledge (ie. the 'intermediate' reason-based approach), has moved our main task from that of coming up with an account of self-knowledge that can accommodate the features of immediacy, first-person authority and immunity to non-cognitive error, to the more specific task of explaining how an experience/ episode of thought/ episode of visualization/ etc. of the *world* (or of some aspect of it) can constitute the direct evidential basis or reason for making a judgement not just about the *world* (or relevant aspect thereof) but also about the mental episode itself (or about some mental disposition one may have towards that same aspect of the world). Put differently, the process of uncovering a satisfactory solution to the traditional problem of authoritative introspective self-knowledge has led us to a deeper puzzle about self-knowledge than the one with which we started.

The first aim of this chapter will be to articulate this puzzle further, and to establish whether, and why, it *must* be addressed head-on. The rest of the chapter will then consider how successful two recent attempts to uphold such an 'intermediate' reason-based view of self-knowledge might be at generating an answer to it.

3.1 Justification and self-knowledge

How can having a first-order (ie. *world*-directed) state itself constitute a direct rational ground, or personal level reason, or direct epistemic basis for believing

something about our *mental states*? In fact, what in the first place is it for a mental state to stand as a reason for believing anything at all – be it about our mental states or about the world?

There are broadly speaking two questions about the epistemic justification of a belief that one could ask, corresponding to two different ways of individuating beliefs: (1) What makes a particular *proposition* believed (or belief *content*) justified; and (2) What makes someone's *believing* that proposition justified. The two are not unrelated.

A *proposition* that p can be said to be justified, roughly, when there is (impersonally speaking) reason to believe it. Someone's *belief* that p, on the other hand, can be said to be justified when this person *has* adequate reasons or grounds for believing it. More precisely, answering a question of the form 'What makes the proposition that p justified?' or 'What reasons (impersonally speaking) are there for believing that p? involves providing an *argument* for p (or pointing to directly suggestive aspects of the world), that is, providing a story (or direct evidence) which entails or makes probable (or directly suggests) that p. For example, answering a question such as 'What makes the proposition that Mary is in the library justified?' or 'What reasons are there (impersonally speaking) for believing that she is in the library?' would involve mentioning, say, (1) that Mary goes to the library almost every day, (2) that she said she was going to be in the library today, (3) that she usually does what she says. Or, alternatively, it might involve just drawing attention (say, by pointing) to (4) her standing in what is visibly a library. In other words, answering a question of the first type involves pointing to available evidence for a particular proposition, that is, to facts, other propositions, ways the world might be or appear which either entail, or make probable, or are directly suggestive of p.

On the other hand, answering a question of the form 'What makes S's *belief* in proposition p justified?' or 'What reasons/grounds does S *have* for believing that p?', involves considering first why the proposition that p is justified, or what reasons (impersonally speaking) there are for believing that p (ie. answering the first question), then restricting one's answer only to those facts (or to those 'reasons' or 'grounds' in the first sense) to which the believer has access. To apply this to our example, answering a question such as 'What makes John's *belief* that Mary is in the library justified?' or 'What reasons/grounds does John *have* for believing that she is in the library' would involve mentioning, for instance, that John heard her say that she

would be in the library and that he believes that she usually does what she says. If, however, John was neither aware of her daily habits nor heard her speak of her plans for the day, or *did* hear her but did not take her to be trustworthy, he would not count as justified in his belief, although in this case the *proposition* that Mary is in the library would still be justified. There would still *be* (impersonally speaking) reason to believe it, although John himself would not *have* reasons or grounds for believing it. Or, put differently, there would still be evidence for p available, though John himself would not *possess* or have *access* to any such evidence.

Thus, in sum, in the first sense, a 'reason' or a 'rational ground' is essentially a fact, proposition or way the world might be. In the second sense, it is a *state of mind* of a particular person. And, someone's having a particular mental state can be said to constitute an immediate reason or ground in the *second* sense for their coming to believe that p, if their having this state somehow involves (perhaps amongst other things) their having *access* (loosely speaking)[1] to evidence for p, that is, their having access to reasons in the *first* sense, ie. to facts, propositions, ways the world might be or appear which either entail, or make probable, or are directly suggestive of p being the case.

Now, given the specific puzzle before us, it is clearly the second sense of 'reason' or 'rational ground' (ie. understood as an internal *psychological* basis for believing something) that is relevant. The puzzling question left to us from chapter 2 was a question about how the transition between two *mental states* can be directly reason-giving, and indeed how it can be reason-giving in an essentially *internalist* sense. We were left, that is, with the need to explain how moving directly from having an attitude towards the *world* to self-ascribing a *mental state* could be something that it 'makes rational sense' to us to do from within our own point of view.[2]

---

[1] Insofar as we are here concerned only with reason-giving transitions between *two* mental states or between an initial set of states and a subsequent belief (ie. independently of what other states one might have, and independently of the origin or grounding of the initial state(s)) this so called 'access' need not be veridical. In the sense relevant here, we can count a person as having reasons for believing that p, or as having 'access' to evidence for p as long as they perceive, or believe, or have experiences *as of* (even if not actually *of*) certain facts, or ways the world might be which entail, or make probable, or are directly suggestive of p.

[2] There are of course other possible accounts available of what constitutes a rational relation, in particular to be found in reliabilist and other externalist accounts of justification. These however seem to go clearly against the phenomenology of belief formation, and more to the point, against the phenomenology of *second-order* belief formation in cases of so called 'introspection' (For our earlier discussion of this, and rejection of purely reliabilist accounts of the relation between our first-order conscious states and our authoritative self-ascriptions of them see chapter 1, sections 1.2.1 and 1.2.2, as well as much of the end of chapter 2). More generally, when making a judgement – be it about the world or about our mental states – we do not seem to just find ourselves, as Evans would put it, 'with a yen to apply some concept' (1982, p.229). Rather, doing so (ie. coming to form a conscious belief – whether first or second-order) is something that makes rational sense to us from within our own *personal-*

So, a mental state is to count, in the sense relevant here, as a reason or rational ground for believing that p, if the content of this state is either directly suggestive of p being the case, or potentially figures as a relevant premise in a possible argument which either entails or makes probable that p, that is, in an argument from which it follows either deductively or inductively that p.[1]

Consider thus again the puzzling question that chapter 2 left us with: how can having a first-order *world*-directed state constitute or give rise to an immediate reason for believing something about our *mental states*? In light of the understanding just laid out of what it is for one mental state to stand as an immediate reason or rational ground for moving on to hold a particular belief, the puzzle implicit in this question (or the explanatory problem faced by the approach to self-knowledge according to which there is a direct reason-giving relation between our first and second-order states) becomes apparent: given that from facts about how the world is nothing either follows, or is directly suggested, about what mental states we are in, how can having a conscious attitude towards the world constitute an immediate reason for believing something about our mental states?

To illustrate this problem, consider first the following simple transition between two initial states (1 & 2) and a subsequent resulting belief (3):

1. John heard (that) Mary say (said) that she would be in the library today
2. John believes that Mary usually does what she says

3. John believes that Mary is in the library

Here, John's reasons or rational grounds for believing that Mary is in the library (ie. states 1 & 2) have as contents the possible premises of an argument or

---

level point of view. We turn to the world for *evidence* or *confirmation* – not just to 'trigger' beliefs – be it about the world or about our mental states.

[1] I am taking it here that perceptual experiences, and not just beliefs, can constitute reasons or rational grounds for beliefs, although of course they cannot constitute *inferential* grounds. It is sometimes held however that beliefs can only be justified by other beliefs because perceptual experiences are not subject to revision; they are not something we are responsible for, and so it does not make sense to say that one *ought* to believe that p if one perceives that p. This however does not seem entirely right, given that we would actually judge someone to be *irrational* if they saw that it was raining but did not believe that it was (assuming of course they had no reason to mistrust their senses). Moreover, given the possibility of disbelief in one's senses, a perceptual experience can surely be taken to constitute an evidential ground, or count as a reason amongst others, for believing something. A perceptual experience is moreover a clear form of 'access' that a believer might have to a justifying fact.

direct pieces of evidence (on the right hand side), which are suggestive of p being the case, or which make it probable that p, or from which it follows (although not strictly in this case) that p. So far so good.

Now, contrast this with another simple case of a mental transition, this time between a first-order state (1) and a *self-ascription* of this first-order state (2):

1. John believes that Mary usually does what she says

2. John believes that John believes that Mary usually does what she says

Clearly, in this case, the conclusion of the possible argument on the right hand side (from the content of 1 to the content of 2) in no way follows from the premise, so how can John's believing the premise constitute an immediate reason from him to move on to believe the conclusion? And, this is just the simplest of examples. Even more mind boggling is the question of how a transition between a world-directed *emotion* or *desire* and a subsequent self-ascription of it could be directly reason-giving in the above sense, since emotions and desires are on the face of it not forms of 'access' to anything at all, let alone to the right types of facts – ie. to facts about our mental states.[1]

There are essentially two ways in which one could try to respond to this problem: (a) one could take it to constitute an actual *problem* for the 'intermediate' reason-based approach to authoritative mental self-ascription, and hence a reason to reject it, or (b) one could take this puzzle as an explanatory *challenge* that must be and can be addressed.

In the next two sections, two such 'intermediate' reason-based accounts of self-knowledge (those of Burge and Peacocke)[2] – ie. accounts according to which there is indeed some form of direct reason-giving relation between our first-order states and our second-order self-ascriptive judgements – will be considered in turn. These would seem to offer us ways of taking line (b). Neither account however will in

[1] To avoid over-complicating things from the outset, the case of our self-ascriptions of emotions and desires will, although touched upon here, returned to in greater detail in chapter 6. The focus in this chapter will be primarily on the puzzle as it arises for cases of transitions between world-directed *cognitive* states and second-order beliefs about them.

[2] Unless otherwise stated, all references in this chapter will be to (Burge 1996). For Peacocke see (Peacocke 1992;1996;1998)

the end be seen to have the resources to fully explain this relation (independently of whether it was their aim to do so), though a close analysis of their discussions will reveal that such a relation *must* hold, and allow us furthermore to uncover the form that a fully explanatory account of this relation will actually have to take. More than that, identifying the limitations of Burge's and Peacocke's approaches to this issue will, we will be see, ultimately narrow down the options available to us for explaining this reason-giving relation between first and second-order states to the extent of revealing exactly how our puzzle is to be solved, and indeed the only way in which it *could* be solved.

More specifically, Burge's account of our entitlement to self-knowledge will be considered first. It will be argued in the next section that, although Burge's account turns out to be unsatisfactory given our present explanatory aim (ie. the aim of explaining how a first-order state can constitute an immediate reason for self-ascribing it), his account can be used to generate a *reductio ad absurdum* of the view that our introspective second-order judgements might *not* be directly and rationally related to our first-order conscious states. An intermediate reason-based approach to self-knowledge, it will thus be suggested, must be adopted, thereby making approach (a) above no longer viable. Approach (b) will therefore *have to* be taken, and the question of how a conscious first-order state can constitute an immediate reason for self-ascribing it will *have to* be addressed. Next, Peacocke's account will be turned to. This account, it will be argued, ultimately faces much the same predicament as Burge's. Its specific limitations however will be seen to narrow down the options even further, to the point of revealing the only possible way in which our puzzle could be solved, and thereby the possibility of immediate authoritative introspective self-knowledge in the end fully accounted for. Let us begin with Burge's account of our entitlement to self-knowledge.

3.2 Burge

According to Burge, the following claims hold:

(1) We have, and are entitled to, a distinctive non-perceptual, epistemically special kind of self-knowledge (by which he means, in the end, knowledge had from and about the same cognitive perspective – understood in the sense defined in chapter 1

above[1] – and so knowledge that stands in an immediate rational relation to that which it is about).

(2) This entitlement to this special kind of self-knowledge arises essentially from the role of this knowledge in critical reasoning.

In support of the first claim, Burge puts forward a transcendental argument to the effect that critical reasoning requires being entitled to a special kind of self-knowledge, and, since we are critical reasoners, it follows that we must be so entitled. More specifically, the argument runs as follows:

(1) We are able to reason critically. That is, we are able to reason in the fully reflective way spelt out in chapter 2 above.[2]

(2) Critical reasoning requires the following:

(i) that we be epistemically entitled to certain judgements about our attitudes and reasons

(ii) that these judgements constitute knowledge (ie. that they be normally true and not just accidentally so)

(iii) that this knowledge be distinctive in being directly rationally related to the attitudes it is about.

(3)Therefore, by (1), we *must* be entitled to this distinctive kind of self-knowledge.

Concerning (i) it is clear that we could not possibly reason critically if we were not entitled to our judgements about our own attitudes, that is, if we were not being reasonable in making them. In fact, if we were not reasonable in our reflective judgements about our own attitudes, we could not be reasonable in our conclusions derived from reasoning based on these judgements, nor could we be reasonable in our reviews of our attitudes based on these conclusions. Critical reasoning would therefore not be possible, given its nature as reasoning which involves *reasonable* confirmation, change, or supplementation of our attitudes based on our reflection on these attitudes. If we were not entitled to our second-order beliefs about our first-order attitudes, this would mean that reviewing or confirming our attitudes on the basis of rational reflection on these attitudes and on our reasons for holding them, would not actually be a reasonable enterprise to engage in, and so, insofar as we are rational, we would not engage in it.

---

[1] See chapter 1, p. 34

[2] See chapter 1, p.36

Concerning (ii), the basic idea is, once again, that if these judgements to which we are entitled did not constitute knowledge, then genuine critical reasoning would not be possible. In fact, we engage in critical reasoning essentially for the purpose of arriving, through reflection on our attitudes, at more reasonable beliefs and at a rationally more coherent set of attitudes. That is, the point of reasoning critically is to *control* or *guide* our beliefs and other attitudes in such a way as to further their reasonability and rational coherence. If, however, our second-order judgements were either never true or only accidentally so, then changing our attitudes on the basis of conclusions derived from reasoning based on these judgements would not be a case of rationally *controlling* or *guiding* our attitudes, even if it somehow accidentally resulted in the promotion of their rational coherence.

Finally, concerning point (iii), as already suggested in chapter 1 by way of an example, critical reasoning seems to require that the knowledge we use in reasoning critically be had from and about the same cognitive perspective, and must therefore be *directly* and *rationally* related to the attitudes it is about. In other words, the idea is that our first-order conscious attitudes must constitute *immediate reasons* for judging that we have them, and similarly, that our second-order evaluative judgements about these first-order attitudes (eg. a judgement that a first-order attitude of ours is unreasonable), must constitute *immediate reasons* for confirming, reviewing or supplementing them.

At this point however, given the explanatory puzzle that this kind of view of the relation between our first and second-order states has given rise to, one might be sceptical about just going along with this argument. In fact, the only kind of argument that could give us a good reason to accept, rather than reject, an intermediate reason-based approach to self-knowledge in light of the clear problem it gives rise to, would be one which could somehow show us that critical reasoning would be *impossible* if it were done from and about a different cognitive perspective. In this way, given that we clearly *are* able to reason critically,[1] our conscious attitudes and our self-ascriptions of them *must* be held from within the same cognitive perspective, and so *must* stand in immediate reason-giving relations to each other.

[1] I am for instance critically reasoning right now in considering and evaluating my beliefs about self-knowledge, with the aim of coming through this process of reasoning at more reasonable beliefs on this topic.

It seems to me that such an argument can indeed be provided, namely the following *reductio ad absurdum* of the view that the immediate self-ascriptive judgements we use in critical reasoning could be held from a *different* cognitive perspective from that of the attitudes they are about.

Let us start by assuming, for the sake of argument, that the self-ascriptive judgements we use in critical reasoning *are* made from a different perspective from the perspective of the attitudes reasoned about, and that they are therefore entirely independent from these attitudes, to which they are related only by a reliable, but purely contingent, non-rational, causal mechanism. Let us then ask whether on these assumptions genuine critical reasoning, as defined in chapter 1 above,[1] would be possible.

For the sake of clarity, let us consider a case in which the relevant dissociation of perspectives is a dissociation between the perspectives of two different people. In fact, let us imagine that my beliefs about your attitudes and your attitudes are linked by some reliable causal mechanism, and, moreover, that I decide to engage in critical reasoning about your attitudes, and begin to reflect on them and on their reasonability, and that I thereby come to the conclusion that you are not being reasonable in one of your beliefs. What would happen next? The process of critical reasoning would seem to be blocked at this stage. The problem is that in critical reasoning, the conclusions arrived at from reasoning based on reflection on one's attitudes, should *immediately* and *rationally* result in an explicit reasonable change or confirmation of the attitudes reasoned about. So, if I were genuinely reasoning critically, then my coming to the conclusion that you were being unreasonable in one of your beliefs should make it immediately rational for you to change your beliefs, which it does not seem to do. In order for my process of reflection on your attitudes to result in your explicitly and reasonably changing your beliefs, this process would somehow first have to result in your coming *yourself* to believe that a belief of yours is unreasonable. This could happen indirectly, for instance, if I were to somehow succeed in convincing you of the unreasonableness of your belief by making you go yourself through the reflection I just went through about your attitudes. In this case however, the reasonable review of your attitudes would end up ultimately resulting from *your* reflection on your own attitudes, and not directly from mine. Alternatively, a *direct* way in which you could

[1] See chapter 1, p.36

come to hold the belief that a certain belief of yours was unreasonable as a result of *my* reflection on your attitudes, would be via some direct (contingent, non-rational) causal mechanism linking the conclusions of my reflections to your beliefs. The problem with this however would be that this causal mechanism could not be the same kind of mechanism by which my second-order beliefs are related to your first-order beliefs, since such a mechanism would only entitle you to the belief that *I* believe that you are being unreasonable in one of your beliefs, but not to the belief that you actually *are* being unreasonable. It would therefore again not immediately follow that there would be reason for you to change your belief. The relevant kind of mechanism would have to be one whereby whatever conclusion I arrived at from reasoning based on my judgements about your attitudes, would immediately, and non-rationally, cause you to hold this *same* belief. Of course now the problem would be that critical reasoning would *still* be impossible, given that such reasoning involves review or confirmation of one's attitudes based on conclusions derived *rationally* from reasoning about one's attitudes, whereas in this case, although *my* conclusion that a certain belief of yours is unreasonable would be directly and rationally derived from reasoning about your attitudes, *your* belief that a belief of yours is unreasonable would *not* be directly rationally derived from any such reasoning.

For critical reasoning to be possible, in other words, all the reflective beliefs about your attitudes as well as the reasoning about them and the conclusions thereby reached would have to be ultimately held by *you*, that is, by the same person whose attitudes are being reasoned about. And, to return to the case of single individuals, this means that in order to reason critically, all our reflective second-order beliefs about our attitudes must be held from the same rational, cognitive perspective as our first-order attitudes.

In sum, our nature as critical reasoners requires that our introspective knowledge of our own conscious states be had from and about the same cognitive perspective as the conscious states thereby known, and, by definition of a cognitive perspective,[1] that there must be a *direct rational* relation between our first and second-order states, in addition to whatever underlying causal relation may or may not also hold between them. In other words, if our introspective self-knowledge were not of this kind we could *know* our own mental states and *reason* about them, but

[1] See chapter 1, p.34

could not thereby immediately rationally influence them, just as, if we had telepathic knowledge of someone else's mental states, this would not give us any kind of immediate rational control over them. But, since in reasoning critically, our reflective judgements *do* immediately rationally influence our first-order states, these judgements must be of the above distinctive kind. Any judgements about our own attitudes which are not based in this direct rational way (such as, for instance, judgements about our own unconscious attitudes), could not be part of a genuine process of critical reasoning. Since, however, the immediate introspective, authoritative judgements we ordinarily make about our states are the very judgements we use in critical reasoning, these judgements must be of this distinctive kind, which however just leads us straight back to our initial problem about how this can be. The only difference now is that the problem is even more pressing, since we can no longer respond to it by simply rejecting the approach to self-knowledge that generates it. The only way forward is thus to try to provide a head-on answer to it. Can such an answer be extracted from Burge's account of our entitlement to self-knowledge?

According to Burge, our entitlement to this special kind of self-knowledge is to be found essentially in our nature as critical reasoners. That is, the role of self-knowledge in critical reasoning is supposed to be the very *source* of our entitlement to it. However, in reading Burge's (1996) paper, one might find that nothing beyond the earlier mentioned transcendental argument is put forward in support of the claim that the role of our introspective judgements in critical reasoning is the *source* of our entitlement to them. Yet, all this argument seems to show is that critical reasoning *presupposes* that we are entitled to a special kind of self-knowledge, but it does not tell us *why* we are so entitled, or how it is that we can be so entitled. One might therefore argue, following Peacocke,[1] that Burge's account gets things the wrong way around. If critical reasoning requires being entitled to self-knowledge, then one must first be entitled to self-knowledge if one is to be able to reason critically. Critical reasoning can therefore not possibly be the *source* of our entitlement but only a consequence of it.

In a sense, Peacocke is right in his objection to Burge. To be fair to Burge however, it is not clear that he is speaking of 'entitlement' or 'source' of entitlement

[1] (Peacocke 1996)

in the same sense as Peacocke. In fact, by way of an example, we can distinguish the following two notions of entitlement:

(1) The sense in which I may, for instance, be entitled to use the concept of a cause (as opposed to just that of constant conjunction) because without it, following Kant's second analogy,[1] objective experience would be impossible. The very existence of experience itself therefore warrants me in my use of it. This, however, says nothing about

(2) What *grounds* I have on a particular occasion for saying that A caused B. What was the basis for my judgement? Did I infer it from some information I already had? Was my judgement based on what I saw? Was it purely a guess? etc. In this sense, I am entitled to a judgement if it is based on, or grounded in, a good reason.

When talking about our entitlement to self-knowledge, Burge seems to have in mind sense (1), whereas Peacocke, in his objection to Burge, has in mind sense (2). That is, for Burge, our nature as critical reasoners warrants or justifies us in our use of our immediate, non-inferential, authoritative, directly reason-based self-ascriptions, in that without them critical reasoning would not be possible. In this sense, the role of our immediate introspective judgements in critical reasoning can indeed be thought of as the *source* of our entitlement to them. Burge is not concerned with what entitles us to self-knowledge in any other sense than this. But, one might ask, why is he not? *Should* he be concerned with explaining what our immediate judgements about our conscious attitudes are based on? Perhaps his thought is that there is nothing illuminating there to be said about what entitles us to self-knowledge in this sense. Our judgements are just immediate; they are not based on any evidence.

As pointed out though already in the introductory chapter,[2] a judgement can be both *immediate* and *rationally grounded.* Moreover, given Burge's account of critical reasoning, it looks like our second-order judgements clearly *are* rationally based on something, namely on our first-order (ie. world-directed) states. As seen above, for critical reasoning to be possible, our first-order attitudes must constitute *immediate reasons* for judging that we have them, and likewise, our second-order evaluative judgements must constitute *immediate reasons* for reviewing or confirming our first-order attitudes. Burge's account in other words does seem to suggest that our

[1] (Kant 1929)

[2] See p.9 above.

introspective second-order beliefs are based on reasons, that these reasons are the very first-order attitudes they are about, and moreover, that these first-order attitudes must count as *good* reasons for our second-order judgements, if critical reasoning is to be a reasonable activity to engage in. Burge's account thus seems to imply that we have a special kind of entitlement to our mental self-ascriptions in sense (2) and not just in sense (1). It does not however *explain* this entitlement, but only leads us more forcefully back to the question set out at the beginning of this chapter, namely that of how our conscious world-oriented attitudes can possibly constitute immediate reasons for our *attitude*-oriented beliefs. That is, how can it be directly rational to move from a conscious attitude towards the world to a belief about oneself and one's mental states?

Burge does not address this question nor does he go into the issue of the rational grounding of our immediate introspective judgements at all, because as mentioned, perhaps he takes it that nothing illuminating can be said about how there can be a direct reason-giving relation between our first and second-order conscious states. All that can be said is that it must be so, because if it were not, we would not be able to reason critically, and it so happens that we *are* able to reason critically. We, however, need to say more than that. Or, at the very least given our explanatory aim, any option in this direction must be fully explored before resorting to the not so obviously coherent idea that the relation between our first and second-order states must necessarily be thought of as *rational* or *reason-giving*, yet in a way which bears no resemblance to the way in which other transitions between mental states of ours are understood as being rational or reason-giving. Perhaps Peacocke has the answer. His account certainly attempts to take things further in the suggested direction than Burge's.

3.3 Peacocke

Focusing on the case of *belief* self-ascriptions,[1] according to Peacocke it is inscribed in the very possession conditions for the concept of belief that anyone who

[1] In his various discussions of introspective self-knowledge, as is common in the literature on the subject, Peacocke focuses primarily on cases of *belief* self-ascription. Presumably however, as will become clear from the account he puts forward of belief self-ascriptions, Peacocke would put forward a similar account of our authoritative self-ascriptions of emotions and desires. We will thus assume his account to extend to such cases.

possesses this concept will find it 'primitively compelling' to judge that they believe that p, whenever they have an occurrent conscious belief that p (and consider the matter), and they will find judging so primitively compelling precisely *because* they have this belief, that is, for the very *reason* that they consciously believe that p. In fact, on this view, one will not count as possessing the concept of belief *unless* one has, in appropriate circumstances, a tendency to make such 'consciously-based self-ascriptions'.[1] In other words, on his view, someone who possesses the concept of belief will make immediate judgements about their occurrent conscious beliefs *directly* and *rationally* on the basis of these very first-order beliefs themselves. Moreover, these self-ascriptive judgements can be said to constitute *knowledge*, or *warranted* true beliefs (and not just true beliefs), because, when made for the very reason that one does have the self-ascribed state, they will be bound to be true. Self-ascribing one's own beliefs in this way can in other words be said to be a reasonable (justified/warranted) process to engage in because it is one which is likely to lead to true beliefs. And, this warrant or entitlement one has to one's 'consciously-based' self-ascriptions can be seen to have its source in the very conditions of what it is to possess the concept of belief, since these conditions are such that anyone who possesses the concept of belief will normally make mental self-ascriptions in the direct way outlined above, that is, directly on the basis of the very conscious beliefs self-ascribed, and so only in circumstances in which they actually do have these beliefs. Given this account of our entitlement to a special kind of self-knowledge, the next question is: how does it compare to Burge's?

In the last section we saw how Burge's transcendental argument from critical reasoning shows that we *must* be entitled to our immediate judgements about our own conscious states based directly and rationally on these conscious states themselves, but it does not actually explain *why* moving from a first-order conscious attitude to a self-ascription of it is a rational or warranted transition; it only suggests that it *must* be. Peacocke's account on the other hand, does, in a sense, seem to explain this: we are *reasonable* in making judgements about our occurrent conscious beliefs on the basis of these occurrent conscious beliefs themselves, because proceeding in this way guarantees the truth of our self-ascriptions. To this extent therefore, Peacocke's account might seem to be an improvement on Burge's. However, the truly

[1] See (Peacocke 1992; 1998)

fundamental question to which Burge's account led us but did not in fact provide an answer was not that of how consciously-based self-ascriptions, given that such self-ascriptions are possible, can generally be veridical (and so how self-ascribing in this manner can be a reasonable enterprise to engage in), but rather, that of how it is possible *at all* for a first-order conscious state to rationally issue in a higher-order belief about itself, given the general understanding laid out at the beginning of this chapter of what it is for a judgement to be reason-based, or of what it is for a transition between two mental states to be, at the personal level, immediately rational.

To put things differently, what Peacocke's account seems to do, is explain how making a second-order judgement directly on the basis of having an occurrent first-order state can be a reasonable (and ultimately knowledge-yielding) process to engage in, and his answer is that this is so because, given the possession conditions for the concept of belief, these second-order judgements will normally be made only in circumstances in which they are true. Engaging in a belief-forming process which is likely to yield *true* beliefs, is clearly a *reasonable* thing to do. This however does not yet address the issue of particular concern to us here, namely that of how this belief-forming process can not only be a reasonable (justified/warranted) process to engage in given its likely outcome, but how it is possible *at all.* We do not, that is, yet have an answer to the question of how a state with one type of content (first-order) can constitute an immediate personal-level reason for making a judgement with a completely unrelated type of content (second-order).[1]

Having said that, it is not entirely true to say that Peacocke does not address this more specific question. There is at least the beginning of an answer to it in his discussion of the 'conscious' character of the beliefs which we are able to self-ascribe in this non-inferential authoritative way.[2]

For Peacocke, a 'conscious' state is essentially a *phenomenally* conscious, or phenomenologically *occurrent* state (in the sense defined in the introductory chapter)[3], or, in Peacocke's terms, a state which is currently 'occupying our attention', and which 'contributes to what, subjectively, it is like for the person who enjoys it'.[4]

---

[1] Martin raises a similar problem for Peacocke in (Martin 1998).

[2] See in particular (Peacocke 1998)

[3] See p.13 above

[4] (1998, p.64)

Now, the relevance that the occupation of attention is supposed to have to our ability to move directly from a conscious state to a self-ascription of it, is that in the case of such self-ascriptions, a second-order judgement is not rationalized by the first-order attitude self-ascribed simply in virtue of this attitude's *content* (in fact, we have seen, how could it?), but essentially in virtue of this attitude's being one which is currently *occupying our attention*, and so somehow contributing to what things are like for us subjectively.

This seems to be a move in the right direction. There must be something about a self-ascribed mental state of ours being *conscious* (or consciously manifested, if it is a dispositional state)[1] that enables it, if anything does, to stand in a direct *reason-giving* relation to a self-ascription of it. This appears in fact to be the *only* aspect of those of our states which we are able to introspectively self-ascribe, that distinguishes them from those of our states which we are *not* able to self-ascribe in any special way (eg. our unconscious or repressed ones). Merely stating this however does not yet resolve our problem – ie. the problem of understanding *how it is* that or *in virtue of what* it is that a first-order state's being conscious, or being such that it contributes to what things are like for us subjectively, enables it to stand as an immediate reason for self-ascribing it. So, how might we go about solving this puzzle?

One possible way of doing this might be by saying that having a phenomenally conscious attitude involves not just having an attitude towards the world, but being also in some sense at the same time *aware* (perhaps implicitly) of oneself as having it. It would thus be in virtue of this self-awareness already present implicitly in having a first-order conscious state, that it immediately 'makes rational sense' to us, when having a conscious first-order state and considering the matter of whether we have it, to judge that we have it. In so judging, we would in effect just be explicitly articulating something we were already aware of implicitly in the having of the first-order conscious state itself. Peacocke however seems to want to resist any such option, particularly as he wants to allow for the possibility that non-human

---

[1] Peacocke does not himself set out to draw a clear distinction between occurrent and non-occurrent (or episodic and dispositional) conscious mental states. He allows in particular that *beliefs* can be conscious in the 'occurrent' sense. It is not clear to me though that beliefs really can be thought of as conscious other than in the Freudian and non-occurrent/dispositional sense. Nonetheless, like other dispositional conscious states, beliefs can it seems come to *manifest themselves* in phenomenally conscious episodes of attention to their objects (eg. in a proposition's coming to strike one as *true* upon being contemplated in thought). I will thus not, for immediate purposes at least, quibble with Peacocke's talk of certain beliefs as being 'conscious' in the phenomenal sense.

animals have conscious states of the very same kind as we do. That is, he wants to leave room for the possibility that animals are roughly the same as us, except that we possess the concept of belief and they do not.[1] In this regard he writes: '...what is involved in a belief's being conscious can be fulfilled by a creature who does not even possess psychological concepts. What is true is that if a thinker does have the concept of belief and has a certain conscious belief, then he will be willing to judge that he has the belief'.[2] In other words: we are capable of self-consciousness and animals are not; and, this is so because we possess psychological concepts and they do not.

But now, if having a conscious state involves, as Peacocke seems to suggest, being aware only of the *world*, we are still left without an answer to our question: how can believing something about the world constitute or give rise to an immediate reason for believing something about our *beliefs* about the world? Perhaps there is something else about a first-order state's being 'conscious' that could enable it to stand as an immediate reason for self-ascribing it, ie. other than the having of it already involving some implicit form of self-awareness. Perhaps, for instance, one could argue that it is in virtue of there being something specific *it is like* to have a conscious (or consciously manifested) belief that p (as opposed to, say, a belief that q, or a desire that z), that this belief can constitute an immediate reason for self-ascribing it. But, how might this suggestion help?

If the proposal is that we base our self-ascriptions of our conscious attitudes on some kind of 'phenomenal feel' that comes with having them (or that comes with conscious manifestations thereof if they are dispositional states)[3], and by which we can somehow 'sense' that we have them, we are just led back to a particularly implausible type of perceptual account of self-knowledge,[4] and to all the problems that go with it. Taking this line is thus not a viable option, nor for that matter one that Peacocke would want to take, given his aim precisely of putting forward an account of self-knowledge that is reason-based *without* being perceptual – an account according to which our introspective mental self-ascriptions are directly and rationally

---

[1] See especially (Peacocke 1992, pp.151-154) and (1998, p.96).

[2] (1992, p.153)

[3] For more on how to understand the 'manifestation' of dispositional states in consciousness see chapter 7 below.

[4] See chapter 1, p.28 above.

grounded in our first-order conscious states *themselves*, and not on some phenomenal feel associated with having them.

On Peacocke's account therefore, given all of the above, the relevance that a first-order state's being 'conscious' (or consciously manifested) might have to its ability to constitute an immediate reason for self-ascribing it, can have nothing to do either with one's being in any sense implicitly *aware* of oneself as having it when having it, or with its being accompanied by some 'phenomenal feel' by which one might be able to sense it. The only answer therefore to our question to be extracted from Peacocke's writings on self-knowledge is, it seems, to be found in his account of what it is to possess the concept of belief, together with his account, derived from his theory of concepts, of what it is for a transition between two mental states to be *rational* or *reason-giving*.

In brief, Peacocke's view is that the transition between a conscious belief and a self-ascription of it is a *rational* transition precisely because it is a transition the making of which is contained in the first-person clause of the possession condition for the concept of belief, namely the following clause: 'A relational concept R is the concept of belief only if [...] the thinker finds the first-person content that he stands in R to p primitively compelling whenever he has the conscious belief that p, and he finds it compelling because he has that conscious belief'.[1] But now, despite Peacocke's promising talk of the relevance of the 'conscious' character of a mental state to it's ability to stand in such an immediate reason-giving relation to a second-order belief about it, it would seem that one can raise a similar objection to Peacocke as Peacocke does to Burge. The above possession condition's being the *right* possession condition for the concept of belief just seems to *presuppose* that we make mental self-ascriptions directly and rationally on the basis of our first-order conscious states themselves but does not *explain* it. We would indeed not count someone as possessing the concept of belief unless they found it primitively compelling to self-ascribe a belief whenever they had that conscious belief (and considered the matter), and found it primitively compelling for that very reason. This however does not yet explain *why* having a conscious belief makes it seem immediately appropriate/rational to us to self-ascribe it, nor does it explain what specifically it is about a self-ascribed

[1] (Peacocke 1992, p.163)

belief's being *conscious* that enables it to stand in such a direct reason-giving relation to a belief about itself.

At this point, Peacocke could of course just insist that a transition between a first-order conscious state and a second-order judgement is rational *precisely because* it is inscribed in the possession condition for the concept of belief, and it is only transitions between *conscious* beliefs and self-ascriptions of them that are inscribed in the relevant first-person clause of the possession condition. In fact to be fair to Peacocke, he seems to have a slightly different view of what it is for a transition to be rational than the one laid out at the outset of this chapter.[1] He simply takes a transition to be a rational one, if it is one that is inscribed in the possession condition for a concept.

For example, a transition from a perceptual experience as of something red, to a judgement that there is something red in the near vicinity, counts as a *rational* transition on his view in virtue of the fact that this transition is written in to the possession condition for the perceptual concept of 'red' – in essence, the condition that someone who possesses the perceptual concept of red will find the content that there is something red in the near vicinity primitively compelling whenever it appears to them that there is something red in the near vicinity and will find it primitively compelling precisely *because* they are having an experience as of something red. Similarly, regarding the concept of entailment, a transition from believing that p and believing that p entails q, to believing that q, is a *rational* transition on his view, because it is amongst the transitions the making of which, or the finding compelling to make which, is written into the possession condition for the logical concept of 'entailment'. That is, in brief, for someone to count as possessing the concept of entailment, they must find, amongst other things, contents of the form q primitively compelling whenever they believe that p and believe that p entails q, and they must find such contents primitively compelling precisely *because*, that is for the very *reason* that, they believe that p and believe that p entails q.[2]

But now, one might still point out that in the case of these other transitions – ie. between these states that are inscribed in the possession conditions for perceptual and logical concepts – we are actually able to go further than Peacocke in making

[1] See section 3.1 above.

[2] All of these possession conditions are laid out in (Peacocke 1992)

sense of *why* someone who possesses the relevant concept (of red or of entailment), and who is in the former state of the transition inscribed in the possession condition, should find the content of the second state of the transition primitively compelling. We can make sense of this in fact by reference to the model of justification set out at the beginning of this chapter, that is, by reference to the idea that they find the latter content primitively compelling in such circumstances because they have *evidence* for the content of the second state of the transition in having the first state, that is, because they have *access* (in having the first state) to facts that *justify* (in the first sense)[1] the content of the second state. To illustrate this, one might represent these rational moves as follows:

(1)
S perceives that p
S believes that p

(2)
S believes that p
S believes that p entails q
S believes that q

(3)
S believes that p
S believes that S believes that p

In cases (1) and (2), we can makes sense of why moving from the first (or first two) mental states of the transition, to the second (or third) state, might be a *rational* transition, or, to put it in Peacocke's terms, why anyone who possesses the perceptual concept of p (or the concept of entailment) and who is in the first (or first two) states, might find the content of the second (or the third) state primitively compelling, by reference to the fact that if the proposition that p (or the propositions that p, and that p entails q) to which they have (loosely speaking) access in having the first states are true, the contents of the second (or third) states of the transition *follow*, that is, will also be true.

[1] See p. 61 above

In the case of the move from having a conscious belief to self-ascribing it on the other hand (ie. case (3)), it remains utterly unclear why someone in the first state of the transition will find the content that they *are* in this state primitively compelling, unless, in having the first-order conscious state of the transition, they either have at the same time some sort of *access* to the fact that they are in this state, or have some sub-personal causal mechanism in their brain which makes them find themselves with a sudden compulsion to self-ascribe a belief whenever they have such a conscious belief. Peacocke however, we have seen, does not want to accept any version of the former option, nor does he want to accept the latter, since, he insists, his account is supposed to be an account of the transition between our first-order conscious beliefs and our second-order self-ascriptions of them as a *personal*-level transition, ie. as a transition which, to use his words 'makes sense to the subject himself given [the subject's] point of view'[1]. This is precisely why he insists that the first-order state in such transitions must be a *conscious* state.

However, until we are given a story about what it is about a state's being conscious that enables it to stand as a personal-level reason for self-ascribing it (other than that it is transitions only between *conscious* states and self-ascriptions of them that are written into the possession conditions for psychological concepts) it remains unclear what sense can be made of this idea that moving from a conscious state to a self-ascription of it is a transition which it makes rational sense to us to make, from within our own personal-level point of view. To take a simple example, what sense are we to make of why, say, from Joe's point of view, his consciously believing that there is a cat on the mat should make him find the content that he *believes* that there is a cat on the mat primitively compelling, if his consciously believing that there is a cat on the mat does not in any way involve his also being *aware* of himself as holding this belief? To put things differently, what could it be about the cat's being on the mat from Joe's point of view that is directly suggestive of, or from which it follows (whether deductively or inductively) that he *believes* that there is a cat on the mat?

Perhaps to be asking these questions is just to be caught up in the assumption that there is more to be said about personal-level justification than a certain transition's being inscribed in the possession conditions for a concept. It is difficult however not to ask them, especially as, first of all, it seems that more *can* be said in

[1] (1998, p.96)

the case of transitions involved in the possession conditions for all concepts other than those of mental states (eg. the concept of red or that of entailment), and secondly, because Peacocke himself stresses the importance of the self-ascribed state's being a *conscious* one if it is to be capable of constituting an immediate reason for self-ascribing it. No real story however of what it is about a state's being conscious that makes it able to constitute such an immediate reason for self-ascribing it is in the end given to us.

In sum, one of Peacocke's main motivations for his account was to be able to maintain, contra what he calls the 'no-reasons' view of self-knowledge, that the transition between our first-order conscious states and our authoritative self-ascriptions of them is a rational *personal*-level transition, that is, a transition which it somehow makes sense to us to make from within our own self-ascribing perspective. However, the only way in which such a transition could seemingly be a *personal*-level rational transition is if that which fixes the content of our self-ascriptions is something to which we are sensitive to, or have *access* to at the personal level, namely something we are *aware* of. Peacocke of course does not deny that making a mental self-ascription involves being sensitive to the psychological nature of our self-ascribed conscious states, since it is in virtue of their psychological nature (rather than their first-order content) that the content of our self-ascriptions of them is going to be fixed. However, we have seen that if this sensitivity is thought of as just a 'phenomenal' sensitivity, we end up with a perceptual model of self-knowledge and all the problems associated with it. On the other hand, if this sensitivity is thought of as a kind of 'implicit self-awareness' intrinsic to the having of a first-order conscious state, then we can no longer hold on to the view (that Peacocke *does* seem want to hold on to) that having a conscious state involves the same thing across species, and in particular awareness only of the *world.* In resisting this latter option though (ie. that having a conscious state involves being at the same time implicitly aware of oneself as having it), his position ends up collapsing either into an implausible perceptual account of self-knowledge, or into a non reason-based view according to which having a psychological state (and not any personal-level sensitivity to it) directly causes one, at the sub-personal level, to self-ascribe it. Taking either of these two routes goes against Peacocke's initial intention of finding an intermediate line between them. He would thus not want to take them. Yet, given his commitments about the nature of our world-directed conscious states (ie. as being attitudes only

towards the *world*), it is difficult to see what other options are left open to him, other than the anti-explanatory one of just saying that there is no further sense to be made of the idea that a transition between a first-order conscious state and a self-ascription of it is rational than mentioning that it is inscribed in the very fact of what it is to possess a psychological concept.

In other words, in the end, Peacocke's account seems to face a similar limitation to Burge's, namely that of failing to provide a satisfactory explanation of how it is that it can be *internally rational* for us to move directly from having a conscious attitude towards an object to believing that we have that attitude towards that object, given that nothing about what attitudes we have seems to follow from how things are with that object. Peacocke's account only seems to point to the fact that we *must* be able to make such self-ascriptions directly on the basis of our object-oriented conscious states, given that doing so is presupposed by the very fact that we possess psychological concepts. He does not however explain *how* this is possible. Nonetheless, his account does seem to move a step further than Burge's in an explanatory direction, first of all in its actually addressing the question, and secondly in its suggesting that it must be something about a first-order state's being *conscious* that enables it to stand as an immediate reason for a second-order belief about it. He seems to fail however to provide a satisfactory story of what this distinctive feature of our conscious states might be.

What point have we now reached? Our examination of Burge's transcendental argument from critical reasoning has first of all shown that our nature as critical reasoners presupposes a certain 'rational integration'[1] of our first-order conscious states and our second-order judgements about them, thereby lending strong support to the view that the 'intermediate' reason-based approach to introspective self-knowledge must indeed be right. Secondly, Peacocke's plausible account of what it is to possess the concept of belief has been seen to also strongly support the view that our immediate introspective judgements about our occurrent conscious states must be directly and rationally based on these conscious states themselves, since we would not tend to regard someone as possessing the concept of belief (or presumably, mutatis mutandis, other psychological concepts) unless they found the content that they believed that p primitively compelling whenever they did consciously believe that p,

[1] (Burge 1996, p.103)

and unless they did so for the very *reason* that they believed that p. Thirdly, as seen in the introductory chapter, the very phenomenology of authoritative mental self-ascription seems also to support the view that we do in fact self-ascribe our own conscious states directly on the basis of having these conscious states themselves. Following Evans's datum on mental self-ascription, when we consider what beliefs (or other conscious states) we have, we do not turn to ourselves or to our mental states for evidence but essentially to the *world.*[1] If I am asked whether I believe that it is raining for example, I will not look at myself but out the window, and then judge on the basis of what I come to see, and on the basis of what I thereby come to believe about the weather, that I either do or I do not believe that it is raining. Finally, if we take this relation between our first and second-order states to indeed be, as suggested, immediate and *rational*, we also have here a clear explanation of why error in self-knowledge is so closely tied to the ascription of rationality, that is, a clear explanation of why introspective self-knowledge is specifically immune to *non-cognitive* error, though not immune to error altogether, an explanation which neither perceptual nor non reason-based views were able to deliver.

In brief, we have seen that there are a number of positive reasons (in addition to the negative ones discussed in chapters 1 and 2) for adopting the 'intermediate' reason-based line of approach to self-knowledge. It is the only approach seemingly able to adequately accommodate all three of the distinctive features of our knowledge of our own minds set out in the introductory chapter. It is also, as seen in the present chapter, presupposed by our practices of critical reasoning and by Peacocke's plausible account of the possession conditions for psychological concepts. However, Burge's intermediate approach to self-knowledge fails to be fully satisfactory given our current explanatory aim, and Peacocke's attempt to fill this explanatory gap by reference to the possession conditions for the concept of belief ultimately faces the same predicament. Nonetheless, the problems raised here for Peacocke's specific proposal, like those raised for Burge's, have not been so general as to undermine the intermediate reason-based line of approach altogether. Rather, the problems encountered have merely revealed that the following three claims, which Peacocke wants to hold together, do not in fact seem to be compatible:

[1] See (Evans 1982, Chapter 7, in particular p.225)

(1) Our immediate authoritative mental self-ascriptions are based on personal-level *reasons*.

(2) Our reasons for these self-ascriptions are the first-order conscious states thereby self-ascribed.

(3) Our first-order conscious states are not in any sense at the same time states of (implicit) self-consciousness; they are attitudes strictly towards the *world*.

Given the arguments of chapter 2 against all forms of non reason-based views of self-knowledge, we have to accept (1), and adopt an epistemological approach to self-knowledge. Given the arguments of chapter 1 (where it was seen that our self-ascriptions cannot be based either on observation of our behaviour or on direct observation of our first-order states), we have to accept (2), and adopt the intermediate reason-based position. The only option left open to us is therefore to reject (3), and to maintain instead that our special, immediate, authoritative knowledge of our own conscious thoughts, beliefs, and other attitudes is based on *reasons*, that these reasons are the very conscious states thereby known, and that such a relation between first and second-order states is possible essentially in virtue of the fact that having a first-order (ie. world-directed) conscious state involves being also (in some sense) at the same time implicitly *aware* of oneself as having it. In other words, if we are to account for the immediate, authoritative, immune to non-cognitive error knowledge we are able to have of the contents of our own conscious minds, we must, it seems, ultimately assume that our world-directed conscious states are already themselves in some sense states of 'implicit', or as I shall sometimes say 'pre-reflective',[1] self-consciousness, a form of self-consciousness that provides the basis for, and comes to be explicitly articulated in, our fully reflective second-order judgements.

[1] I am borrowing this term from (Sartre 1969) though I will be using it here (unless otherwise specified) with no presumption to be conforming exactly to Sartre's own use of it.

Chapter 4: Consciousness as self-consciousness

What does it mean to say that our first-order conscious states are states of implicit or 'pre-reflective' *self-consciousness*? What is it for the having of a world-oriented state to involve not only awareness of the world but also of itself? Our examination of the various available theoretical lines of approach to introspective self-knowledge has revealed that some such claim must be true, that is, that if that our world-directed conscious states are to constitute the immediate rational grounds for our second-order judgements (which we have seen they *must* if immediate authoritative self-knowledge is to be possible) we must assume them to also be in some sense states of implicit or 'pre-reflective' self-consciousness. This leaves us however without an entirely clear sense of what the claim is supposed to amount to, and in particular without a clear sense of how an account of introspective self-knowledge as based on the nature of our conscious states as states of 'implicit' or 'pre-reflective' self-consciousness is supposed to differ from the many accounts of self-knowledge already on offer.

Spelling out this thesis more clearly will be the main task of this chapter. We arrived at it, to repeat, essentially through seeing that some such claim *must* be true if self-knowledge, of a kind that exhibits certain distinctive features, is to be possible. The best way therefore to begin to understand what exactly the view is to be taken to amount to, is by looking at what it *must* amount to if it is indeed to constitute the view it is supposed to constitute, namely one that enables us to account for the possibility of immediate, authoritative and immune to non-cognitive error self-knowledge in a way that avoids the many problems facing other available accounts. Bearing this in mind then, this chapter will begin by first spelling out what the thesis should *not* be

taken to be, before going over what it *has* to involve, and finally by considering what further positive sense might be made of it. First then, what is the thesis *not* to be taken to be?

4.1 Perceptual accounts, constitutive accounts, and the higher-order-thought theory of consciousness

First, and perhaps most obviously, the required form of 'implicit' self-consciousness should *not* be taken to be a form of accompanying *perceptual* awareness of ourselves as being in this or that state. In particular, the suggestion that having a conscious mental state involves at the same time being implicitly aware of oneself as having it should not be taken to be the suggestion that our conscious states have some kind of phenomenal buzz associated with them by which we are able to sense them, as this would just lead us straight back to the many problems already seen to be associated with perceptual accounts of self-knowledge, and in particular with one of the most implausible versions of it. This, of course, is not to deny that our occurrent conscious states are states which there is something 'it is like' for us to be in, or to deny that we are somehow sensitive to them via their phenomenal properties. The only point to remember here is that 'what it is like' properties associated with having particular mental states, and the accompanying sensitivity we may have to our conscious states via these phenomenal properties, is not sufficient for the grounding of the distinctive kind of fine-grained mental self-ascriptions we are clearly able to make in introspection, and so cannot be taken to constitute the crucial 'implicit self-awareness' that must be involved in the having of a first-order conscious state if our puzzle is to be solved. The sense in other words in which the first-order states that occur in our phenomenal stream of consciousness must be at the same time states of implicit self-consciousness *cannot* be the sense in which we might be sensitive to them via their phenomenal properties.

Secondly, the 'pre-reflective' form of self-consciousness being suggested here to be involved in consciously experiencing/ thinking about/ visualising/ etc. the world, should also *not* be taken to be a version of the 'strong constitutive' or 'artefact of grammar' approach to self-knowledge discussed in chapter 2. Saying that the possibility of introspective self-knowledge presupposes that our conscious world-directed states are intrinsically states of implicit self-consciousness, should not be

taken to be the view that it is somehow *ontologically constitutive* of having a first-order conscious state that one also has a corresponding higher-order belief about it or vice versa. That is, the proposed thesis should not be taken to be the view that our first-order states and second-order beliefs about them are in some sense one and the same state, or the view that our first and second-order judgements are both *expressions* of one and the same state, or for that matter any form of the view that the mere having of a first-order state (or the making of a second-order judgement) somehow ontologically constitutes in one the having of a second-order belief about it (or the first-order state self-ascribed). To make any one of these claims would just be to call forth the many problems already seen to be associated with strong constitutive accounts of self-knowledge, and in particular to bring back the problem common to them all, of their being unable to adequately accommodate the *fallibility* of self-knowledge, that is, to accommodate its immunity to *non-cognitive* error without ruling out the possibility of error altogether.

Finally, the thesis that our phenomenally conscious states are states of implicit self-consciousness should not be merely taken to be the thesis that our conscious states are states which are always *accompanied by* second-order beliefs about themselves, since, even if this were true, merely stating this fact would not explain *why* it is so, or how it *can* be so. To make such a suggestion would not be to offer a solution to our puzzle at all, but instead to give rise to the initial problem of introspective self-knowledge all over again, ie. that of explaining *how* or *in virtue of what* our conscious states are such that we always know (or are immediately *able* to know) that we have them. What we need to identify here is a form of self-consciousness that can provide the direct *ground* for a second-order belief about it, but which is not itself already a second-order belief.

Of course, one could argue following 'higher-order thought theories of consciousness' such as Rosenthal's,[1] that a conscious state's being accompanied by a non-conscious second-order belief about itself is just a primitive fact about our phenomenally conscious states, a fact which cannot be analysed or explained any further. It is not an additional form of 'self-knowledge' based on our first-order states, but simply something that systematically comes with the having of conscious states. According to that theory, a state is conscious just in case it is accompanied by a non-

[1] See (Rosenthal 1991)

conscious second-order belief about itself, and its being conscious indeed *consists* in its being so accompanied. A number of objections however can be raised against this kind of approach to consciousness, not least of which the simple consideration that we can, and often do, have second-order beliefs about attitudes of ours which are *not* conscious, thus revealing that a state's being accompanied by a higher-order thought about itself is clearly not a sufficient sign of its being conscious, let alone the very fact that makes it conscious. I may for instance discover through psychoanalysis that I have feelings of resentment towards a member of my family, and thereby come to hold the second-order belief that I have these feelings. This however will not necessarily make my feelings of resentment become conscious. They could well remain completely repressed, to the point that I may even start doubting whether what my analyst led me to believe is true. Moreover, without having to appeal to repressed attitudes, one can imagine having both a first and a second-order belief, both of them playing an active role in affecting one's actions and other attitudes without *either* of these beliefs actually coming to occupy one's attention, that is, without either of them becoming conscious in the phenomenal sense. I may for instance both believe that my name is Isabella and believe that I *believe* that my name is Isabella (as might be exhibited in various ways in my behaviour), without either of these contents (ie. that my name is Isabella, or that I believe that it is) for long periods of time actually *occurring* to me, that is, without my actually *attending* to these contents, or without their coming to consciously strike me as 'true'. In other words, the 'higher-order thought' theory of consciousness seems to be neither plausible as it stands nor *a fortiori* a theory which could be taken to provide the fundamental primitive fact that underlies the possibility of introspective self-knowledge.

In brief, we have seen in the course of chapters 1-3 that if the original problem of authoritative introspective self-knowledge is to be solved, our first-order conscious states and our second-order self-ascriptive judgements about them must satisfy a number of criteria with respect to each other. (a) They must be *distinct* states from each other. (b) The latter must be *rationally based* on the former (not just causally triggered by them or based on any intermediate observation of them). And importantly, as seen in chapter 3, (c) this must be so essentially in virtue of the intrinsically *self-conscious* character of the former. In light of the points made here moreover, this new required 'self-conscious' character of our first-order states cannot itself be taken to consist in any one of the forms (or bases) of self-knowledge already

seen to be problematic. It can consist neither in our being somehow *perceptually* sensitive to our first-order conscious states when having them, nor in their not being entirely ontologically distinct states from our reflective second-order judgements about them after all, nor of course, given our present explanatory aim, in their having a characteristic which would just generate the same explanatory question all over again – such as in their being merely *accompanied* by non-conscious second-order beliefs about themselves, or indeed in their being the direct rational grounds for our self-ascriptions of them. But, if this is what being a state of 'pre-reflective' self-consciousness *cannot* not be taken to amount to, what *can* it be taken to amount to? What else is there for it to possibly be? Where are we to even turn to for further insights?

4.2 Pre-reflective self-consciousness, the threat of infinite regress, and the phenomenology of world-directed consciousness

An appealing further suggestion can it seems be found in Sartre's discussion of self-knowledge in *Being and Nothingness*.[1] The exact form of self-consciousness that we need in order to solve our puzzle about the grounding of our second-order judgements directly by our first-order conscious states is in fact, I believe, very close to what Sartre might have had in mind when speaking of the 'pre-reflective cogito' as a form of self-awareness implicit already in consciousness itself, which is then made explicit in the fully reflective Cartesian cogito, for which it stands as the latter's 'pre-cognitive' basis.[2] Now, in accordance with our own conclusions, this suggestion seems to present reflective (ie. fully explicit) self-knowledge as being based on, and as being only possible in virtue of, the presence of an implicit, non-articulated form of self-consciousness already in the very act of having the world-directed conscious states we are able to introspectively self-ascribe.

Sartre's discussion however of this 'pre-reflective' or 'pre-cognitive' form of self-consciousness remains somewhat limited. It stresses at length (as we have come to do in this thesis) the importance of the existence of a *pre-reflective* form of self-consciousness for the possibility of fully *reflective* introspective self-knowledge, but

[1] See (Sartre 1969)

[2] See (*ibid*, especially pp.xxvi-xxvii, and chapter 2, section III)

does not elaborate on it much further. In particular, Sartre's discussion still leaves us very much in the dark (a) about the precise nature of this required 'pre-reflective' form of self-consciousness, (b) about how this notion might *concretely* be seen to apply, if at all, to our ordinary world-directed experiences, thoughts, emotions, etc., and (c) about how appealing to a 'pre-reflective' form of self-consciousness to solve our puzzle about introspective self-knowledge might not just amount to turning the problem of explaining how introspective *self-knowledge* is possible into that of explaining how an implicit form of *self-consciousness* is possible, thus in effect not so much solving our puzzle as moving it elsewhere. In other words, although we have in Sartre's proposal (unlike in some of the other proposals considered above) the very thing we are looking for – ie. a form of self-consciousness present already at the world-directed level which can in turn provide the needed basis for a direct rational move from a first-order conscious state to a second-order judgement – we do not yet have in it a substantive account of what it *is* for a first-order state to be a state of 'pre-reflective' self-consciousness, nor any immediately obvious indication about where to look for what it might be. Sartre merely arrives at the conclusion that there *must* be such a thing.

A similar thought to Sartre's can also be found implicit in Kant's point about it having to be at least possible for the 'I think' to accompany all our representations,[1] suggesting, somewhat similarly, that the grounds for the possibility of fully reflective self-knowledge are present already in our world-directed experiences themselves (ie. in the special character of our 'representations'). As it stands though, this claim too leaves us without a concrete sense of how these grounds for the possibility of fully reflective self-knowledge could already be present in our way of experiencing the world. How are we to move forward from here? Let us look at this in context.

To be basing our second-order judgements directly on our first-order (ie. world-directed) states is, we have seen, essentially to be coming to make these judgements somehow not by 'looking inside', but by looking (or focusing our attention in thought, imagination, etc.) *outward* onto the world. This is of course what gave rise to our puzzle about explaining how being in possession of evidence regarding the world, or how in some cases just attending to the world, could possibly directly and non-inferentially, yet at the *personal* level (ie. in an internal

[1] See (Kant 1929, p.153 or B132)

epistemological sense) provide us with an immediate basis for believing something about our own *attitudes* towards it. Put in the form of an example, the question was this: what could it be, say, about a cat's *being* on the mat from our point of view, that immediately suggests to us that we *believe* that there is a cat on the mat, or, even more puzzlingly, that we *fear* the cat, or want to stroke it? Given this way of putting the question, it seems that it must be something precisely about the way the *cat* strikes us in experiencing it or in consciously contemplating it in thought, memory, imagination, etc. that immediately suggests to us that we have a particular attitude towards it. The obvious way forward would thus seem to be to examine the phenomenology of *world*-directed consciousness (ie. to examine how the cat actually *does* strike us in experience, thought, imagination, etc.) in order to see how this could be, *rather than* to look for some further account of 'self-consciousness' in terms of which the 'implicit' form of self-consciousness we are trying to gain some insight into could be understood.

To an extent, this direction of inquiry is hinted at both in Sartre's' discussion of the 'pre-reflective cogito' and in Kant's point about our 'representations' presupposing self-consciousness. Sartre's discussion in particular, in stressing that fully reflective self-knowledge is just an articulation of what is already there in *world*-directed consciousness, suggests that it is indeed to the nature of *world*-directed consciousness that we should turn for further insights rather than to some further explicit account of self-consciousness or self-knowledge. Similarly, Kant's point about our representations being always potentially accompanied by the 'I think' though not always having it explicitly attached to them, suggests that the road to fully understanding our ability to attach the 'I think' to all our representations may ultimately lie in a better understanding of the nature of our representations, that is, in a better understanding of the way in which we consciously experience, think about, visualise, etc. the *world*. And indeed, given the specific puzzle before us, turning our attention to the way in which we experience the world may well be the only viable way forward. To persist in trying to articulate the needed 'implicit' or 'pre-reflective' form of self-consciousness by way of yet other explicit notions of self-consciousness or by appeal to accompanying second-order beliefs or accompanying second-order perceptual experiences, will only make our explanatory quest for a substantive account of this implicit form of self-consciousness impossible to complete. Put more precisely, the problem is this:

Trying to articulate what a particular form of self-consciousness is, by saying that it is a particular (however implicit or 'pre-cognitive') form of 'consciousness of oneself as being in this or that state', immediately establishes a reflective dissociation between the form of self-consciousness to be explicated and that which it is consciousness *of*. Any such dissociation in turn gives rise to the question of how this new form of consciousness of our own mental states is possible, and of what *it* is based on. If one then follows the line followed so far in this thesis, one ends up having to rule it out as being either based on inference or on observation or on 'nothing', and having to conclude that it is based on yet a more fundamental form of self-consciousness present already even more 'implicitly' (if this even makes sense) in the state it is consciousness *of*. Ultimately, one is just sent off on an infinite regress of having to explain, each time at a different level, how self-consciousness of a more and more fundamental or implicit kind is possible, and of having to do this by positing ever more fundamental forms of self-consciousness, which when in turn articulated (ie. said to be forms of 'awareness of oneself as being in this or that state') immediately give rise to the same question all over again of how *these* forms of self-awareness are possible, and of what *they* are based on, etc.

In light of this threat of infinite regress, two options seem open to us: (1) to conclude in despair that the 'pre-reflective' form of self-consciousness needed to solve our puzzle about mental self-ascription cannot be articulated beyond a certain point; it must at some point be taken as *primitive*, and yet as something that *must* be present in first-order consciousness as the ground for its possible reflective articulation in a second-order judgement; or, alternatively (2) to turn away, as suggested above, from thinking explicitly about *self-consciousness* and the different forms that this might take (ie. some perceptual form, the form of a mere accompanying second-order belief, or the form of some special feature of a certain category of *statements*), and to turn instead to thinking about the nature, or rather the phenomenology, of *world*-directed consciousness – ie. to thinking about the way in which the world strikes us, from within our own personal-level point of view (note, regardless of how it actually *is*) in experience, thought, imagination, etc. such that it might contain, amongst other things, evidence already of ourselves as having this or that cognitive or emotional attitude towards it. If any fully *explanatory* solution to our puzzle is to be found it will clearly have to be found through option (2).

4.3 Solving the explanatory puzzle

Our investigation has once again taken a new turn, this time from an enquiry into the nature of an implicit form of self-consciousness to an enquiry into the nature of *world*-directed consciousness. Let us first briefly recapitulate how we got here.

We began this thesis with the traditional problem of explaining how immediate, authoritative, immune to non-cognitive error knowledge of our own minds is possible. We were then led, through an examination of the specific shortcomings of all standard approaches to this problem (inferential models, perceptual models, and non reason-based accounts) to the view that if such knowledge (which we *do* seem to have) is to be possible, we must ultimately assume that it is, in a personal-level sense, based on *reasons*, not however either on inference or observation, but directly on the conscious states thereby known. This conclusion in turn, combined with a clearer understanding of what it is for a transition between two mental states to be directly reason-giving at the personal-level (discussed in chapter 3), gave rise to a new and deeper explanatory puzzle: that of explaining how a first-order conscious state can possibly constitute an immediate rational ground for self-ascribing it, given that our first-order states are about the *world* and our self-ascriptive judgements about our *mental states*, and given that nothing about what mental states we are in directly follows from (or is directly indicated/expressed by) how things are in the world. To put things differently, we ended up in chapter 3 with the problem of having to explain why someone (or, at least a fully rational and conceptually competent person) should find it immediately appropriate to apply a mental concept to themselves (say, that of perceiving), in a situation in which it would appear that they are in no way aware of the fact that the concept is applicable (ie. of the fact that they are having a perceptual experience), since by hypothesis, the situation is one where they are looking only out at the world and not also at themselves and their states. Confronted with this new puzzle, and with the specific limitations (from an explanatory point of view) of two recent attempts (Burge's and Peacocke's) to uphold such a direct reason-based view of self-knowledge, we eventually had to conclude, at the end of the last chapter, that in looking out at the world we must be aware not only of the world (though that is what we are primarily attending to), but somehow also implicitly or 'pre-reflectively' of ourselves as *perceiving* the world, or as having some other underlying dispositional attitude towards it.

Yet, this left us with a further explanatory problem: this time, that of explaining how, concretely, attending to the world can indeed involve being aware not only of the world but also of oneself as having a particular attitude towards it. The conclusion reached so far in the present chapter has been that if any truly substantive insight is to be gained into how our second-order abilities might be already reflected in the way we experience, think of, and visualise the world, we must turn away from explicitly thinking about *self-knowledge* and *self-consciousness*, and turn instead to thinking about the way in which we actually *do* experience, think of, and visualise the world. That is, the focus of our enquiry must now turn away from considerations about self-consciousness and turn instead towards direct considerations about the phenomenology of *world*-directed consciousness. The remainder of this thesis will do just that.

Before embarking on this seemingly vast new task though, a number of preliminary points should be borne clearly in mind regarding what exactly it is that, in the next few chapters, we will and will not need to uncover about world-directed consciousness.

(1) First, it should be noted that, for our specific purposes, we will *not* need to establish that being conscious or having conscious states *consists in* one's being at the same time implicitly or 'pre-reflectively' self-conscious, or that being pre-reflectively self-conscious in having a first-order state is in any way *essential* to this state's being a distinctively conscious one. We can in fact readily allow, in agreement with Peacocke for instance, that what it takes to have a *conscious* mental state is something that can be fulfilled even by a creature who does not possess any psychological concepts, and who is perhaps not capable of any form of self-consciousness at all – whether of a conceptually articulated form, or of some 'non-conceptual' form.[1]

What we *will* need to do, is to uncover a (possibly contingent) feature of *our* first-order conscious states which could explain how *we* (even if not other creatures) are able to move directly from consciously attending to the objects of our attitudes to self-ascribing these attitudes towards these objects. In other words, the task ahead of us is a somewhat narrower one than it might initially appear. Having to show that/how the conscious states which we are able to introspectively self-ascribe are at the same

[1] I will discuss the possibility of the existence of 'non-conceptual' forms of self-consciousness in the next chapter, noting in particular how such 'non-conceptual' or 'pre-conceptual' forms of self-consciousness (as portrayed in the recent literature – eg. Bermudez 1998) should be distinguished from the implicit or 'pre-*reflective*' form of self-consciousness needed here to solve our puzzle about mental self-ascription.

time states of implicit self-consciousness is *not* to have to show that self-consciousness is somehow constitutive of, or presupposed by, consciousness in general.

(2) A second, and closely related point to stress before we proceed any further is that given the above narrower remit of the claim to be articulated, success in our endeavour to show how *our* ordinary thoughts, experiences, etc. might indeed be states of 'pre-reflective' self-consciousness, should not be taken to have many of the counter-intuitive consequences associated with accounts of consciousness as *itself* presupposing self-consciousness. In particular, the claim arrived at here should not, as mentioned above, be taken to have the undesirable consequence that non self-conscious creatures cannot be said to have conscious states of any kind at all.

Having said that, one might still feel that our (albeit restricted) thesis has an equally counter-intuitive result. To say that *our* conscious states are states of 'implicit' or 'pre-reflective' self-consciousness does after all rule out non-human animals and very young children (whom, I am assuming, we do not take to be self-conscious) from qualifying as having conscious states of the same kind as *we* have (even if not from qualifying as having conscious states altogether). And, it is perhaps *this* consequence that we find most counter-intuitive, even when considering more general claims about consciousness as presupposing self-consciousness. This point is worth looking into further, as it tends to constitute a primary reason for wanting to avoid theories which entail that having conscious states (of any kind) requires or presupposes being (in any way) self-conscious. Let us therefore briefly consider how threatening this point really is to our findings.

Upon reflection, it is not I believe as damaging to the intuitive appeal of our position as it might initially seem. For one thing, most of us would agree that neither non-human animals nor very young children have occurrent conscious *thoughts* of any kind, let alone thoughts of the same kind as our own. We would not, that is, intuitively be inclined to say that animals and young children explicitly *think to themselves* – be it in words or in images – that is, that they engage in conscious acts of entertaining possibilities, in conscious acts of judgement (ie. acts of assent to propositions), or indeed in any episodic mental acts requiring the explicit conscious manipulation of symbols with meanings. To deny that non self-conscious creatures have conscious *thoughts* is of course not to deny that they may have many *non*-conscious (ie. non-occurrent) *dispositional* states such as beliefs, desires and

emotions, which guide their behaviour and which are clearly manifest in their behavioural patterns. The view arrived at in this investigation is only that *consciously* (in the *occurrent* sense) attending to the objects of our attitudes involves being pre-reflectively aware of ourselves as having them. If the dispositional attitudes of a creature are thus not manifest in any such occurrent conscious acts of attention, no pre-reflective form of self-consciousness need be involved in having them. It is only derivatively, through the manifestation of some of *our* dispositional states in occurrent conscious acts of attending to their objects (ie. in thinking of their objects, perceiving them, etc.) that we are able to come to know our own dispositional mental states.[1]

Now, insofar as we would not be prepared to think of non-human animals as engaging in any occurrent conscious acts of *thinking*, their dispositional states can of course not come to manifest themselves in any such conscious acts (ie. conscious thoughts). If any real clash is therefore to be seen to arise between the conclusions of our investigation so far, and our intuitions regarding the subjectivity of non-human animals, it will have to be seen to arise essentially in relation to considerations about the conscious *experiences* of non-human animals – ie. states which we would generally take them to have, and which, we may feel, must be similar to our own. The question we therefore need to ask ourselves is this: would we really want to say, from an intuitive standpoint, that the conscious experiences (or episodes of sensory recollection, of visualisation in dreaming, etc.) of non-human animals and very young children are, phenomenologically, of the very same kind as our own?

It is far from obvious it seems to me that we *would* want to say this. For example, it is far from clear that we would want to say that non-human animals experience the world as an objective, unified world, of usable objects, etc. in the way that we seem to do, rather than as McDowell for instance suggests, as a series of obstacles, opportunities, problems and other pressures from the environment, not experienced as such, but merely dealt with as they come.[2] Consider for instance whether we would really want to say that bats experience the world in the same way as we do (or rather, in the same way as we would, if we had echolocatory experiences). Would we want to say that dogs experience the world in the same way as we do, or primates? In the case of bats we would probably be inclined to say that

[1] This was discussed in the introductory chapter section 0.2 above.

[2] See (McDowell 1994, chapter 6)

they do not, while in the case of dogs we might be more hesitant, and perhaps in the case of primates we might even be tempted to say that some do. It is noteworthy though that, in each case, the degree to which we would be prepared to say that a creature's conscious experiences of the world are similar to our own corresponds roughly to the degree to which we would be prepared to attribute to this creature, amongst other things, some level of self-consciousness. At the very least therefore it remains far from obvious that there is anything deeply counterintuitive about the view that the way the world appears to non self-conscious creatures might be, phenomenologically (ie. from a personal level point of view), quite different from the way it appears to us. It would be highly surprising in fact to find that both self-conscious and non self-conscious creatures alike, conceptually competent and non conceptually competent creatures alike, humans as well as bats, etc. all experienced the world in the same way, the only real difference between species lying in the fact that some but not others possess additional conceptual skills and additional faculties such as of 'introspection'. If we *did* discover that all animals experienced the world in exactly the same way as we do, and that therefore, for instance, the way in which fish subjectively experience their environment is just as we would experience it if we were in a fish's body swimming under water, would we not be inclined to say that these animals were to an extent *self-conscious*? If we would, this would suggest that our reluctance to attribute self-consciousness to non-human animals actually reflects an underlying intuitive reluctance to think of them as subjectively experiencing the world in the same way as we do. And, if so, clues to our special ability to self-ascribe our own mental states directly on the basis of consciously attending to their objects may well be there to be found in the distinctive character of the way in which we in particular experience these objects.

(3) Finally, and before proceeding to try to make more concrete sense of this thesis about our own conscious world-directed states, one further issue is worth addressing concerning it still in the abstract – ie. the issue of the *status* of this position. How should it really be taken? Should it, in particular, just be viewed as a 'default' position, one which we have reason to accept (and for the moment no obvious reason to reject) only because no better route to solving our initial problem of introspective self-knowledge seems available? Or, can it be seen to be a plausible view even on positive independent grounds? There is I believe already much evidence to suggest the latter.

Although we arrived at the view that our authoritative mental self-ascriptions are based on our first-order conscious states considered as states of 'pre-reflective' self-consciousness essentially through a process of elimination of alternative approaches to the problem of introspective self-knowledge, the position reached here, upon reflection, is not only able to avoid the many problems faced by its counterparts, but also ultimately the one that seems to fit best with the phenomenology of the process of introspection discussed in chapter 1.

First, if our phenomenally conscious states are indeed intrinsically states of implicit self-consciousness, we can now immediately see why (as discussed in chapter 1 section 1.2.1) self-ascribing them might make immediate rational sense to us from within our own outward-looking and self-ascribing point of view. Next, if our thesis is correct, we also have an explanation of why (as also discussed in 1.2.1) when it suddenly occurs to us that we are, say, thinking about something (or angry at someone, etc.), this information does not usually strike us as a surprise, but is something that we tend to feel we were aware of all along – even if we were not explicitly thinking about the fact that we were in this or that state. Finally, and relatedly, if it is indeed true that in consciously attending to the world we are in some sense at the same time pre-reflectively aware of ourselves as having this or that attitude towards it, we also have an explanation of why (as also discussed in 1.2.1) we are often able to *remember*, much later, thinking thoughts, or having other attitudes (beliefs, desires, emotions, and so on) which we were not at the time reflecting on. For example, we now have an explanation of why I may be able to remember *thinking* to myself as a child that Santa Claus does not exist (thus putting me now in a position to self-ascribe this thought), although at the time I was only thinking about the existence or not of Santa Claus and not about myself and my thoughts.

In sum, the solution to the problem of introspective self-knowledge reached in this thesis by way of the assumption that our first-order conscious states are at the same time states of implicit self-consciousness, appears to be, in the abstract at least, not only *necessitated* by the shortcomings of all other possible theoretical approaches to introspective self-knowledge, and not only *unperturbed* by our intuitions about the subjectivity of non-human animals, but arguably also in many *positive* ways the most intuitively plausible.

What we need to do now is to try to make more concrete sense of this thesis. We have seen in this chapter that the only way forward towards a substantive account

of how experiencing, visualising, etc. the world might indeed involve our being implicitly aware of ourselves as having a particular attitude towards it, is through a direct enquiry into how we actually *do* ordinarily experience, think of and visualise the world in having particular cognitive or emotional attitudes towards it. The time has come to take on this task.

Chapter 5: Cognitive states: self-consciousness, experience and objectivity

So: how *do* we experience, think of and visualise the world at the first-order level? How does the world ordinarily strike us when we consciously attend to it? And, in particular, how is it that this way the world strikes us can provide us with information simultaneously about the world and in so doing (ie. not in some separate act) also somehow of ourselves as perceiving it, thinking of it, or indeed as having some other underlying dispositional cognitive or emotional attitude towards it? The aim of the remaining chapters will be to consider how evidence of ourselves as having particular *cognitive* attitudes towards the world on the one hand (in this chapter), and how evidence of ourselves as having particular *emotional* attitudes towards it on the other (in chapters 6-8) might indeed, in practice, be seen to be already present in the way the world appears to us/ is for us phenomenologically, in conscious experience, thought, memory, imagination, etc.

Starting in this chapter with the case of self-ascriptions of cognitive states, the immediate question before us is this: how does the world ordinarily strike us from within our own outward-looking point of view (regardless of how things might actually *be*) in experience, thought, imagination, memory, visualisation, etc. such as to be immediately suggestive to us of our having a particular *cognitive* perspective upon it, that is, of ourselves as either perceiving it, thinking about it, or holding a particular belief about some aspect of it?

An appealing answer to this question, this chapter will argue, can be extracted from Strawson's interpretation of Kant.

5.1 Self-consciousness and experiencing the world as objective

That there is something peculiar about the phenomenology of world-directed consciousness that can explain the possibility of fully reflective introspective self-knowledge is, it was suggested in the last chapter, hinted at both in Sartre's writings on the 'pre-reflective cogito' and in Kant's point about it having to be at least possible for the 'I think' to accompany all our representations. Sartre's discussion of the 'pre-reflective cogito' in *Being and Nothingness* did not however tell us anything sufficiently specific about this way the world actually is (or appears, or is thought about, etc.) from our point of view such as to potentially reveal to us something simultaneously about the world and about our having a subjective perspective upon it.[1] A somewhat more informative account however of how our world-directed conscious experiences might be at the same time states of implicit self-consciousness can I believe be extracted, albeit in a modified form, from Kant's discussion of self-consciousness, and in particular from the connection Kant draws between the possibility of attaching the 'I think' to all our representations on the one hand and our experiencing the world as an *objective* world on the other.[2] This Kantian idea, as it might apply to our purposes, is most clearly articulated by Strawson.

Strawson's suggestion is essentially that experiencing the world as an *objective* world can be said to have a dual aspect, in that taking the order and content of a series of such experiences (ie. experiences as of an objective world) together (say, as articulated in a series of judgements), would give us 'on the one hand a (partial) description of an objective world and on the other a chart of a single subjective experience of that world. Not only the series as a whole, but each member of the series has a double aspect'.[3] In other words, to experience and think about the world as an *objective* world is, in effect, to experience that which our experiences are as of as not *exhausting* the world (ie. as not constituting the *whole* world but only *part* of a world which extends beyond that which is presented to us), and to experience the order in which the world presents itself as not necessarily being the order in which things *exist*, which, in effect, just is to experience that which our experiences are as

[1] See again (Sartre 1969, pp.xxvi-xxvii, and part two, chapter 2, section III)

[2] See (Kant 1929, the Transcendental Deduction (B))

[3] (Strawson 1966, pp.105-106)

of, as being the world only *as experienced from our point of view* or *as thought about by us*.

In relation to our present concerns, it thus seems that insofar as we *do* experience the world as objective (even if we may not necessarily *believe* it to be objective)[1], experiencing the world in the way that we do can be seen to involve making an implicit distinction between how things are on the one hand and how things are experienced as being from our point of view on the other, and so, in effect, to involve our having not only experiences as of a *world*, but also, in so doing, as of ourselves as having a particular cognitive or perceptual point of view upon it. In other words, as required to solve our puzzle about mental self-ascription and thereby our initial problem about introspective self-knowledge, experiencing the world in the way that we do does seem to involve being somehow implicitly or as I am calling it 'pre-reflectively' self-conscious. More than that, if it is right to say that we do experience the world as objective, and that doing so does involve drawing an implicit objective/ subjective distinction of the kind described by Strawson, it turns out not only that consciously experiencing the world in the way that we (adult human beings with first and second-order abilities at least) do involves being pre-reflectively self-conscious, but that it cannot *but* involve being pre-reflectively self-conscious. We could not experience the world in the way that we do (ie. as objective, as independent from our having a perspective upon it) without being self-conscious. In this Kantian account of objective experience as having a dual aspect we may thus finally have a positive account of how our second-order abilities might already be reflected in the very way in which we experience (or think about, visualise in imagination etc.) the *world*; and, moreover, an account which does not just replace the problem of explaining how a special form of *self-knowledge* is possible with that of explaining how a special form of *self-consciousness* is possible, but an account that actually explains and makes concrete sense of how an implicit form of self-consciousness can already be present in world-directed consciousness itself.

It should be noted however that neither Kant nor Strawson have anything like our puzzle about mental self-ascription in mind when discussing the connection

[1] To say that we experience (or indeed think about, visualise etc.) the world as objective is not to say that our experiences, etc. are necessarily *of* an objective world or even of a world *believed* to be objective, but merely to say that *phenomenologically* and prior to reflection our experiences (episodes of thought, imagination, etc.) are (to repeat) *as of* an objective world. The metaphysical issue of what our world-directed conscious states are actually of or believed to be of – ie. real objects, sense data, or nothing at all – is not of immediate relevance to our purposes, though it is of some importance to Kant's. More on this below.

between self-consciousness and the objectivity of our experiences. Kant (and Strawson in his interpretation of Kant) is not so much concerned with how we might be able to self-ascribe our own mental states directly on the basis of looking out at the world, as keen to establish that the world of our experiences must be objective, or at least conceived of as being objective. Kant's starting point is the fact that we are self-conscious, that is, that we are always capable of attaching the 'I think' to all our representations (ie. of attributing our successive experiences to ourselves, to a single unified mind or perspective). He then argues from this that our experiences must be of a world conceived of as being objective (ie. mind-independent and spatio-temporally connected), on the grounds that only experiences of a world conceived of as such can make room, in Kant's terms, for the possibility of always attaching the 'I think' (the same 'I think', expressing the unity of our experiences) to all our experiences. In brief, his aim is to show that self-consciousness presupposes experiencing an objective world (or a world conceived of as being objective), and he does this, at least on Strawson's interpretation, roughly by arguing that if self-consciousness is to be possible, our world-directed experiences or 'representations' must be such as to be immediately attributable to ourselves, and it so happens that only experiencing a world conceived of as being objective can make them so attributable; our experiences must therefore be of a world conceived of as being objective.[1]

For our purposes, in appealing to Kant's and Strawson's discussion of the connection between self-consciousness and the objectivity of our experiences we are clearly not (nor need we be) following the dialectic of the above Kantian transcendental argument. For the purposes of explaining how it can make immediate rational sense to us, at the personal-level, to move directly from attending to the objects of our attitudes to self-ascribing certain attitudes towards these objects, we do not need to establish that the world of our experiences either *is* objective, or *conceived of* by us to be objective, let alone that our experiences *must be* (even if only phenomenologically) as of an objective world. For our purposes, it only matters that phenomenologically the world does *in fact* strike us as objective, whether or not it *must* do so by virtue of any requirement of self-consciousness.[2] And, on the face of it,

---

[1] (Strawson 1966)

[2] Recall, in the first part of this thesis it was established that for self-knowledge to be possible, our world-directed conscious states must be states of implicit or 'pre-reflective' self-consciousness. It was *not* established, nor is it being argued here, that the *only* way in which they could be so is by being as of an objective world. The only point being extracted here from Strawson's interpretation of Kant is that being phenomenologically as of an objective

our ordinary experiences *are* as of an objective world. More than that, even our hallucinatory experiences, visualisations, recollections, etc. seem to be as of an objective world in the relevant sense.

Consider for instance visualising a tree in imagination. In visualising a tree in imagination, one is not taking any actual objective tree to be before one, or even to exist. Yet, in so imagining a tree, what one appears to be doing is, phenomenologically at least, not projecting or conjuring up a mind-*dependent* tree (eg. some kind of 'sense datum' of a tree – whatever that might look like) but an *objective* tree, a tree presented from a particular angle (as opposed to, say, presented from 'every angle' or from 'no angle' at all – if this even makes sense to consider), and thus, in effect, a tree *as viewed by one* from a particular point of view. The phenomenology of imagination too in other words, and not just of perception, seems to involve drawing the kind of objective/ subjective distinction that allows for the possibility of moving directly from attending to a world (or imagined world) to self-ascribing a cognitive or experiential (whether veridical or not) perspective upon it. Much the same could also be said of the phenomenology of hallucination or indeed experiential recollection. What we appear to be presented with in hallucination, as what we conjure up in imagination, or have in mind in recollection, is not a mind-dependent world but an *objective* world, a world always presented from a particular point of view, and this, it seems, regardless of what, if anything, is actually metaphysically before us, or conceived by us to be metaphysically before us.

A number of metaphysical theories are compatible with this phenomenology of first-order consciousness. Staying with the example of imagination, one could argue for instance that in imagining a tree one is in fact confronted with a sense datum. Or, one could argue that one is not actually confronted with anything at all – it only appears that way phenomenologically. Alternatively, or additionally, one could argue that in imagining a tree one is in a representational state with (perhaps amongst other things) *propositional* content, or, on the other hand, in a representational state with purely *imagistic* content. Following some of the recent psychological literature, one could argue further that in visualising something in imagination, a representation that is essentially imagistic in nature actually comes to be formed in our brain (as suggested for instance by the fact that much the same parts of our brain come to be

---

world is at the very least *one* way in which a series of world-directed states can be seen to be states of 'pre-reflective' self-consciousness.

used in visualisation as in ordinary perception),[1] and manipulated in essentially 'visual' type ways (eg. by way of essentially *spatial* transformations being performed in our heads, etc.) when we engage in reasoning from this representation – when, say, trying to determine whether a piece of furniture will fit in a particular corner of our living room, without actually measuring anything.

Which if any of such theories of the metaphysics of imagination is correct, or which if any of these theories is correct about the metaphysics of hallucination or of ordinary perception remains however essentially irrelevant to our purposes. For the purposes of solving our specific puzzle about mental self-ascription, what matters is, as mentioned above, not the *metaphysics* of imagination, hallucination, ordinary perception, etc. but the *phenomenology* of it, and in particular the phenomenological fact that our experiences, visualisations, hallucinatory experiences, all seem to be as of an *objective* world.[2] If a mind-independent table is before us for instance and being perceived by us, but does not actually strike us as such phenomenologically (if it appears to us instead, say, in the confused way in which it appears to newborn infants or certain non-human animals – not as an object amongst others, arrayed in space, presented from a particular angle, etc.), we will not, in perceiving this (albeit objective) table be making the kind of objective/ subjective distinction that comes with experiencing the world as objective, and which could be appealed to to explain how attending to the objects of our first-order attitudes can provide us with direct grounds, from within our own point of view, for self-ascribing these attitudes. What is important in other words, for self-knowledge to potentially arise out of our experience of this table, is not that the table actually *be* objective, but that it strike us as such phenomenologically.

In sum, we do not, for our purposes, need to refute idealism or establish that the world of our experiences is believed by us to be objective. We also do not need to establish, following a Kantian line, that the world of our experiences *must* (even if just phenomenologically) present itself to us as objective – there may be other ways in

---

[1] See for instance the discussion of this in (Kosslyn 1995)

[2] In saying that our experiences, hallucinations, and other world-directed states are phenomenologically as of an objective world, I am of course *not* saying that there are *no* phenomenological differences between these states. I am also *not* suggesting that what is metaphysically before us in having such states might not in certain respects make a *phenomenological* difference to what is involved in having them. The sole point being made here is that these states have in common the phenomenological characteristic of being as of an objective world – a characteristic which they can have whether or not they are actually *of* an objective world, or believed to be of such a world.

which our first-order states, or those of other creatures, could be states of pre-reflective self-consciousness. Our starting point here is not (nor need it be) the possibility of self-consciousness as something from which to derive the objectivity of our experiences. Instead, our starting point is merely a plausible assumption about the actual (even if perhaps not in principle necessary) phenomenology of world-directed consciousness, an assumption that can it seems be appealed to to show that and how our first-order conscious states can make room for immediate thought about themselves (essentially along the lines described by Strawson), thus allowing us to make concrete, non-metaphorical sense of how consciously thinking about, experiencing, recalling, imagining, etc. the world can contain already within itself the grounds for the possibility of fully reflective introspective self-knowledge – the latter (ie. reflective introspective self-knowledge) just being an explicit articulation of what is already there implicitly and intrinsically (ie. not as an additional accompanying second-order state) at the world-directed level.

Having said this, solving the puzzle in the above way (ie. by reference to the phenomenological objectivity of our experiences) can of course only truly succeed if the assumptions made in so doing can be legitimately taken on board – ie. the assumption that we do ordinarily experience (think of, visualise, etc.) the world as objective, and the assumption that experiencing the world in this way does in fact involve drawing an objective/ subjective distinction of the kind described by Strawson. Moreover, even granting these two assumptions, a number of further questions might be raised concerning this solution to our puzzle more generally:

One might wonder for instance how appealing to the objectivity of our experiences in order to explain the possibility of self-consciousness can avoid being essentially circular, given that experiencing the world as objective and being implicitly self-conscious are being said here to come essentially *together*. Put differently, how, one might wonder, can the possibility of self-consciousness be truly explained by an explanandum (ie. the phenomenological objectivity of our experiences) which is being said to already presuppose its explanans (ie. the existence of an implicit form of self-consciousness)?

Next, one might be concerned that even if this appearance of circularity can somehow be dispelled, we might be faced with an explanatory regress instead. Will we not, that is, in explaining the possibility of self-consciousness by way of the objectivity of our experiences end up at best solving the explanatory problem of how

an implicit form of *self-consciousness* is possible by bringing forth a new and equally puzzling explanatory problem this time about how experiencing the world as *objective* is possible?

Third, even assuming that these concerns about circularity and regress can both be dealt with, one might still wonder what exactly is to be understood by the 'implicit' or 'pre-reflective' character of the self/world dualism supposedly involved in experiencing the world as objective. To say that a self/world distinction is being drawn 'implicitly' could it seems mean a number of things (eg. non-conceptually? without attending to what one is doing? etc.) not all of which – as we will see in what follows – may actually serve the purpose of solving our specific puzzle.

Finally, and perhaps most pressingly, even if the objectivity of our experiences can successfully be appealed to to solve our puzzle about mental self-ascription for the case of *cognitive* states and perceptual experiences, a question still remains about how our experiencing the world as objective can possibly help explain our ability to self-ascribe a wide range of our own *desires* and *emotions* towards the world. How is it, that is, that the objectivity of our experiences is to be appealed to to help make sense of how we are in many cases able to move directly from attending to the objects of our emotions and desires (eg. a cup of coffee that we desire, a spider that we fear) to judging that we would *like* a cup of coffee or that we are *afraid* of spiders? In brief, how are we to fit mental states other than judgements, thoughts, and perceptual experiences as of an objective world into the picture so far given of how a state can be a state of 'pre-reflective self-consciousness'?

The rest of this chapter will attempt to address these many questions in turn, leaving the last one however (ie. concerning the self-ascription of desires and emotions) to be pursued in greater detail in the subsequent chapters.

5.2 Objectivity and self/world dualism

First then, why should we believe that we experience the world as objective and next, why should we be persuaded that thinking of, or experiencing the world as objective (assuming that we do so) has the kind of self/world dual aspect described by Strawson and needed to solve our puzzle about mental self-ascription?

It has been assumed so far in this chapter that our world-directed experiences, memories, hallucinatory experiences, etc. are phenomenologically at least *as of*

objective objects, events, states of affairs, etc. – eg. as of a red dragon 'out there' before us (eg. in hallucination), as of a table viewed from above (eg. in imagination), as of an event unfolding and continuing to unfold independently of our having a perspective upon it, etc. Even our first-order *thoughts*, it is being assumed here, are (unreflectively at least) as of objective tables and chairs, as of objective events and states of affairs, as of *possible* objective states of affairs, and so on. We think and talk, that is, as of mind-independent objects and events, whether or not we believe the things we are talking about or thinking about to be in actual fact mind-independent.

Some people may feel however, when they reflect upon the phenomenology of their experiences in particular, that the world strikes them as mind-*dependent*, although they may on occasion come to *believe* or judge this world to be, contrary to the way it presents itself to them, an objective world. Matters of phenomenology are by their very nature difficult if not impossible to establish against contrary intuitions. It will thus to a great extent just have to be assumed here, on the grounds of (on my view) greater intuitive plausibility, that the world does not strike us as mind-dependent but as objective, and this thesis taken as addressing itself only to those who would agree that we ordinarily experience the world as objective. In the context of discussing the phenomenology of *our* world-directed conscious states however (ie. world-directed conscious states of the kind that we, adult human beings with sophisticated first and second-order abilities have), this is an assumption that most people would probably find plausible. No attempt will in any case be made here to argue, say, by way of a Kantian transcendental argument, that our experiences *must* be phenomenologically as of an objective world. It may be that such an argument can be given, but doing so here would lead us to a far stronger conclusion than we currently need. For present purposes, all that needs to be granted is that our experiences *are* phenomenologically as of an objective world, whether or not they *need* to be so by virtue of any special feature of self-consciousness or of the unity of consciousness. In taking the phenomenological objectivity of our experiences on board however merely as an empirical assumption, the possibility ought to be considered, even if only briefly, that this assumption might be misguided, and that we might in fact experience the world as mind-dependent. How threatening would our experiencing the world as mind-dependent really be to our attempt here to show that the way in which we experience the world involves being pre-reflectively self-conscious?

If someone were to insist that we experience the world as mind-dependent, or that at least this is how the world strikes *them* phenomenologically, it could be quickly pointed out that no real difficulty need arise from this for the claim that experiencing the world in the way that we do involves being implicitly self-conscious. Suggesting that we experience the world as mind-dependent would in fact make our task (of explaining how experiencing the world in the way that we do involves being pre-reflectively self-conscious), if anything, easier. Consider what it would actually be to be experiencing the world as mind-dependent. Consider in particular what it might be to be having an experience as of a sense-datum, or what it would be to be having a hallucinatory experience by which one is not taken in (in the very act of experience, not just in judgement). The very idea of a *sense*-datum or of a hallucinatory object is that of a *subjective* entity, in the sense of a *mind-dependent* entity, ie. of something that exists only insofar as *it is perceived.* In experiencing the objects of one's attitudes therefore *as* sense-data, hallucinatory objects and so on, reference (however implicit and non-articulated) to oneself as having a cognitive or experiential perspective upon these objects is going to be inevitably smuggled in. One could not be having an experience *as of* a mind-dependent entity without thereby taking oneself (implicitly, if not downright explicitly) to be in some mental state or other, and so without tending to find it appropriate, upon asking oneself the question of whether one is having an experience, to judge that one is. The view that we experience the objects of our attitudes as mind-dependent can in other words quite easily be fitted into the picture of self-knowledge put forward here so far.

In sum, whether one agrees that we experience the world as objective, or wants to claim instead that we experience it as mind-dependent, experiencing the world in the way that we do phenomenologically can be seen to involve being at the same time aware of ourselves as being in some mental state or other. It will continue to be assumed here nevertheless that in most cases we experience the world as objective (ie. as mind-*in*dependent). This, on my view at least, is both the more intuitive assumption to make and the more interesting one, in that experiencing the world as objective can be seen to involve our being simultaneously aware of the world and of ourselves as having a perspective upon it *despite* the fact that at first glance at least (unlike experiencing the world as mind-dependent), experiencing something as

objective involves making reference only to the thing attended to and not also to oneself and to one's mental states.[1]

Leaving thus aside the suggestion that we might experience the world as mind-dependent, one distinctive feature of the world-directed conscious states that we at least have (ie. adult human beings with sophisticated first and second-order abilities) has now been identified, a feature that can be seen to enable us to immediately and authoritatively self-ascribe a cognitive attitude simply on the basis of attending to the object of this attitude. This feature is the phenomenological *objectivity*, from our point of view, of the objects of our world-directed attitudes.

Taking this on board, let us move on next to consider the second assumption made in this chapter: the assumption that experiencing the world as objective involves, as suggested by Strawson, drawing an objective/ subjective distinction. In brief, I take this to be a straightforward conceptual point. To experience the world as objective in the present sense just *is* to experience it as independent from our perspective upon it, or, put differently, as independent from it's being *present* to us. Again therefore, no argument will be put forward to show that this *must* be the case, as this is just how I understand and will continue to understand the notion of the 'objective' in this context, ie. as that of something 'independent of one's having a perspective upon it'. To experience the world in a way which *did not* involve drawing such an objective/ subjective distinction would not be to experience it as objective in the way that we do. A conception or direct experience of the world as objective (in the present sense) is, I take it, intrinsically a self/ world dual conception or experience.[2]

---

[1] A concern might of course remain about whether the world can actually be said to phenomenologically strike us *as* anything at all – be it as mind-dependent or mind-*in*dependent. Perhaps, prior to any concept being applied in an act of judgement the world does not strike us as anything at all, except in the confused way in which it is often claimed to strike non-human animals, ie. as just a series of obstacles, not experienced as such, but just dealt with as they come. This seems both phenomenologically highly counterintuitive (in our own case at least), and is something which would lead to a whole host of further problems about how we might be able to move from such experiences to the kinds of judgements we arrive at based on our experiences. Moreover, it might be claimed following Bermudez for instance (see Bermudez 1998, ch.8) that even non-human animals and very young children (let alone adult human beings with a sophisticated understanding of the world) experience the world *as* objective, although somehow 'non-conceptually'. Issues about conceptual and non-conceptual content will be returned to in more detail in the next section. In the meantime, it will continue to be assumed here, as it seems most plausible to do, that we experience the world *as* something, and in particular *as objective*, whether or not explicit concepts should be thought of as being deployed in our so doing.

[2] There are of course other notions of the 'objective' and corresponding notions of the 'subjective' that may be appealed to in different contexts. For an illuminating discussion of three possible ways of drawing the objective/ subjective distinction see for instance (Eilan 1997). I am assuming that it is relatively clear from what has been under discussion, that the notion of the 'objective' in play here is one which applies to the *world* as experienced (rather than to the character of our representations thereof), and, that the distinction between the objective and the subjective is a distinction between the mind-independent world before us and our having a point of view upon it.

Moreover, it is worth noting that, aside from this point appearing obvious to me, it has been taken to be so in much of the literature. Strawson for instance, when discussing (in an unrelated paper) whether a subject of a non-spatial (ie. purely auditory) experience could be a subject of an objective experience, takes the question 'Can the conditions of knowledge of objective particulars be fulfilled for a purely auditory experience?' to mean 'Could a being whose experience was purely auditory, make use of the distinction between himself and his states on the one hand, and something not himself, or a state of himself on the other?'[1] That objective experience involves drawing a distinction between one's states on the one hand and something not oneself or one's states on the other is taken for granted. Similarly Evans, in his commentary on this discussion of Strawson's takes it for granted that objectivity and the drawing of a self/world distinction come together, the substantial issue for him also being that of whether conceiving of the world as objective (and thereby drawing a subjective/objective distinction) presupposes conceiving of it as spatial.[2] Another question that arises in the literature about the presuppositions of objectivity in the present sense is that of whether being able to have objective experience and to draw the distinction it involves presupposes being, or at some point having been, an *agent* and thereby having experienced resistance to one's will.[3] Again, these discussions about the connection between objectivity and agency take for granted that objective experience and the drawing of a subjective/objective distinction come together, the real issue being whether being able to do so requires being an agent.

In other words, although the intimate connection I am appealing to here between objectivity and self/world dualism may not have been put to the particular use to which I am trying to put it here (ie. that of providing a grounding for introspective self-knowledge), it has certainly been noted, and quite rightly it seems to me, been assumed to hold. To experience the world as objective in the way that we do just *is* to experience it as independent from the particular perspective we have upon it.

To sum up then, experiencing the world as objective just is to experience it as distinct from one's perspective upon it, and this essentially because a conception of it as objective, when made fully explicit, just *is* a conception of it as distinct from one's

---

[1] (Strawson 1959, p 69)

[2] (Evans 1985)

[3] See for instance (Baldwin 1998) and (Russell 1998) in *The Body and the Self.* eds. Bermudez, Marcel and Eilan.

perspective upon it. That these two conceptions are inextricably linked allows us moreover to make sense of why it seems impossible for instance that someone could possess sophisticated first-order concepts of the kind that we have and be in a position to make explicit first-order judgements (eg. of the form 'There is a book on the table'), yet *not* be in a position to make any second-order judgements (eg. of the form 'I *believe* that there is a book on the table' or 'It *seems* to me that there is a book on the table') due to their not happening to possess the concepts of belief or experience. If what has been said so far is anything along the right lines, such a situation could not occur, since one could not possess first-order concepts of the kind that we have without possessing some second-order concepts, and so could not be in a first-order conscious state which involved the deployment of these first-order concepts, without being in any position at all (whether immediately or otherwise) to self-ascribe them. To put things differently, the suggestion being made here is that the concepts of experience and belief are not something that one might (or might not) acquire *after* having acquired first-order concepts, and similarly that having second-order abilities is not an *additional* faculty to that of having sophisticated first-order abilities. The two come together. Thinking about the world or indeed directly experiencing it (imagining it, etc.) as an objective world, and thinking about it or directly experiencing it as an *experienced* or *thought about* world, are just two sides of our same intrinsically self/world dual way of thinking about or experiencing the world.

One might be concerned nevertheless that an inevitable consequence of this is that only beings who possess some notion of the 'subjective' can be taken to experience the world as 'objective'. What about those creatures (eg. primates perhaps and young children) whom we may want to think of as experiencing the world as objective (following some of the recent psychological and philosophical literature for instance)[1], yet whom we would not want to think of as possessing any notion of subjectivity?

As discussed in the previous chapter,[2] it seems to me that we would actually be quite reluctant to attribute grasp of the world as objective to a creature whom we did not also think of as being to some extent self-conscious, and so as having some

[1] Gibson and Bermudez amongst others argue that some non concept possessing creatures exhibit behaviour which suggests that they distinguish in experience between the world and their subjective perspective upon it. See the discussion of Gibson and Bermudez further below in section 5.4

[2] See 4.3 point (2) above.

grasp of the idea of their having a subjective perspective upon the world. Of course, we would probably not want to think of primates or very young children as possessing any fully fledged concept of 'subjectivity' or as being capable of fully reflective self-knowledge, but neither would we, I take it, want to think of them as having a fully fledged concept of 'objectivity' or as being capable of forming explicit thoughts about the world before them. The kind of grasp, if any, of the objectivity of the world that non-human animals and very young children might be taken to have (assuming that they do not possess either first or second-order concepts) is going to have to be in some sense *non*-conceptual or *pre*-conceptual. This is in fact the kind of grasp attributed to them in the literature mentioned above. It is attributed to them moreover on the basis of evidence that suggests that, in experiencing the world, they are drawing some form of distinction between the world before them and their point of view upon it, a distinction which they must presumably also be drawing *non-conceptually* since they do not possess any second-order concepts.[1]

One can it seems therefore accept that some creatures experience the world as objective *without* having to deny the point made in this chapter that experiencing the world as objective and experiencing oneself as having a perspective upon it come inextricably together. The form (conceptual or non-conceptual) of self-consciousness that comes with experiencing the world as objective, will just in each case correspond to the form (conceptual or non-conceptual) of a creature's grasp of the world as objective.[2] Nothing in fact about what has been said so far suggests that grasping the world as objective in experience (and drawing the corresponding objective/ subjective distinction) must be an all-or-nothing affair or a fully conceptual affair. If one wants to leave room for the existence of less sophisticated conceptions of the world and of its objectivity, or for non-conceptual experiences of it as objective, one can just say that the *degree* to which one's conception or grasp of the world's objectivity in experiencing it is sophisticated will correspond to the *degree* to which one's conception or grasp of one's own subjectivity is sophisticated, and in turn to the degree of sophistication to which one can be thought of as *self-conscious* in having this experience.

---

[1] That non-conceptual forms of self-consciousness are possible, and characteristic of the consciousness of some animals and young children, is something argued for strongly by Bermudez in (Bermudez 1998).

[2] A creature's grasp of the world's objectivity may go anywhere from being a purely 'practical' grasp (cf. Campbell 1994, ch.1, p.30), to perhaps being a 'non-conceptual' representation of it as objective, all the way to being a fully conceptual, complex, theoretical, multifaceted understanding of it as objective.

Having said this, nothing essential to the progress of the present thesis hangs on whether one is or is not prepared to allow for degrees of self-consciousness or for non-conceptual forms of first and second-order awareness. As long as experiencing the world as objective in beings like ourselves capable of fully explicit reflective self-knowledge comes together with awareness of ourselves as experiencing it, our solution to the problem of introspective self-knowledge stands unperturbed. Whatever the case of non concept-possessing creatures may be, consciousness and self-consciousness can be seen to come inextricably linked in us, since we could not experience the world as objective in the way that we actually do (given what this involves, ie. drawing an objective/ subjective distinction) without experiencing ourselves as having a cognitive or perceptual point of view upon it. It is worth reminding ourselves in fact that the whole purpose of introducing the notion of pre-reflective self-consciousness in this thesis was to allow us to answer just this one very specific question of how, in an adult human being like ourselves, a direct, rational, personal level move, on a particular occasion, is possible from a conscious thought or experience to a judgement about this thought or experience. For this narrow purpose a number of things are and can be taken for granted.

First, it can be assumed that we, those for whom the problem of self-knowledge is being raised, *are* able to have first-order conscious thoughts and experiences of the kind that we actually have, namely ones which (we have agreed) are phenomenologically as of an objective, spatio-temporal world, as of a world of usable objects (eg. tables and chairs), as of objects that are presented from a particular perspective and so available for demonstrative reference (eg. objects experienced as being *there* or *this* one) and so on. Next, it can also be assumed that we possess all the concepts (if any) that might be required for having such first-order states, together with whatever else this might presuppose. It can also be taken for granted that we possess all the *second-order* concepts required for making explicit mental self-ascriptions, since our whole problem arose from the fact that we are able to self-ascribe our own mental states in a seemingly special way. That we *are* able to make mental self-ascriptions is not being put into question. We clearly are capable of fully explicit self-knowledge. Creatures who do not possess second-order concepts may well turn out to be self-conscious in some way, but there is no need to assume here that the form of self-knowledge that *we* have, or the form of self-consciousness that is implicit in *our* experiencing the world as objective, is going to be of the exact same

kind as that of those creatures. In our case, the relevant 'implicit' form of self-consciousness could, for instance, be *conceptual* in some sense.[1] The crucial point here is that the problem of introspective self-knowledge and puzzle about self-consciousness to which it leads arise *even* once it is assumed that we possess first and second-order concepts, and *even* once it is assumed that we have world-directed experiences as of an objective world. The problem remains that, in self-ascribing our world-directed experiences and other mental states, we still attend only to the *world* (even if indeed grasped as an objective world composed of physical objects arrayed in space, etc.), and on *that* basis make judgements about our *mental states*. It is the possibility of *this* move that the notion of pre-reflective self-consciousness, and then considerations about the objectivity of our experiences, are supposed to explain. We need not therefore show that, in attending to the world, the implicit form of self-consciousness involved in our doing so is not of a kind that presupposes possessing second-order concepts, since we *do* possess second-order concepts, and since our problem arises even once this is assumed.

Of course, none of this tells us anything about how possessing second-order concepts or having a sophisticated conception of the world as mind-independent of the kind that we have is possible in the first place. Thinking however of self-consciousness and of the phenomenological objectivity of experience as not being an all-or-nothing affair (as suggested a paragraph back) may make it easier to see how our fully-fledged form of self-consciousness and our sophisticated, fully conceptual, theoretical grasp of the world's objectivity and of our own subjectivity might have progressively developed from some more primitive earlier situation through our interaction primarily with the *world*, and through our coming to construct a progressively more complete and elaborate picture of this world. The difference between us and other animals and young children may lie not in us, as opposed to them, being self-conscious at all, or having any grasp of the world as objective at all, but rather, in our having an overall more sophisticated conception of the world and of its objectivity, with our sophisticated second-order abilities and sophisticated

[1] The fact that we possess sophisticated first and second-order concepts does not of course entail that the form of self-consciousness that is implicit in our experiencing the world as objective is of a kind that requires possessing second-order concepts. It could be of the kind that non concept-possessing creatures can also have, ie. of a kind that does *not* require possessing second-order concepts. The point here is that, for our purposes, we do not *have to* show that this implicit form of self-consciousness is in some sense 'non-conceptual' (ie. of the latter kind) since we (for whom the problem is being raised) *do* have second-order concepts.

conception of ourselves as having a perspective upon the world just being a particularly distinctive sign of this.

A lot more could of course be said about the exact connection that exists between the kind of conception we have of the world and of it's independence from us on the one hand and the kind of conception we have of ourselves and of our subjectivity on the other than I will be going into here, but, given all that has been said, it does seem that this is going to be a very intimate one. In the case of other animals too, we are likely to find that the kind of grasp a being has of itself and of it's subjectivity is reflected in the kind of grasp it has of the world and of its objectivity. Any concerns in any case related to the kind of understanding that non concept-possessing creatures may have of the objectivity of the world before them and of their own subjectivity need not worry us here. We clearly have a sophisticated grasp of both. The puzzle of mental self-ascription for us starts from the fact that *we* (whatever the case may be of other animals) are able to self-ascribe our own mental states upon merely attending to the world. This is a puzzle that arises for fully rational, conceptually competent (at both first and second-order levels) adult human beings with experiences, thoughts and other attitudes of the very kind that we have. What has been suggested here is that this puzzle can be solved by reference to the fact that we experience the world as objective, and by drawing attention to the fact that experiencing the world as objective involves drawing a distinction between the world before us and our particular perspective upon it.

A number of points of clarification remain however to be made. Recall in particular some of the further questions raised at the end of the last section.

First, how can appealing to the phenomenological objectivity of our experiences possibly help explain the possibility of self-consciousness by virtue of the fact that this phenomenological objectivity *presupposes* self-consciousness? Does this not involve some form of explanatory circularity? Next, if it does not, are we not led down the road of explanatory regress instead? Are we not, that is, in explaining the possibility of self-consciousness by appealing to the phenomenological objectivity of our experiences at best just shifting our central question of how an implicit form of *self-consciousness* is possible to that of how experiencing the world as *objective* is possible? Let us consider these two worries in turn.

5.3 Circularity and explanatory regress

A key point made in this chapter was that experiencing the world as objective and drawing a distinction between the objective and the subjective always come together, that is, that experiencing the world as objective and being implicitly self-conscious are just two sides of the same coin, two sides of our same intrinsically self/world dual way of experiencing the world. The solution put forward here to our puzzle was then that it is precisely *in virtue* of this fact (ie. of the fact that objectivity and self-consciousness come inextricably together) that we are able to self-ascribe our own conscious states directly on the basis of having them, ie. directly on the basis of attending to their objects out in the world. Such accusations therefore as that objectivity already presupposes self-consciousness (or, in Kant's terms, that experiencing the world as objective already presupposes the possibility of attaching the 'I think' to all our representations) and so cannot be used to explain it, are based on a misunderstanding of the solution here on offer and of the puzzle to which it is a solution. Far from being viciously circular, the explanation of how being implicitly self-conscious at the world-directed level is possible by reference to the fact that we experience the world as objective, is an explanation that can (a) help *clarify* the sense in which objectivity can be concretely said (following Sartre, Kant and Strawson) to already 'presuppose' self-consciousness or to 'make room for thought about itself' and (b) provide an account of *why* it is always possible (following Kant) to attach the 'I think' to all our conscious 'representations', or (in less Kantian terms) to immediately self-ascribe our own mental states. It is always possible because having the self-ascribed mental states themselves, insofar as having them involves experiencing the world as objective, also involves thereby being implicitly aware of ourselves as having them. The fact therefore that the existence of a pre-reflective form of self-consciousness is *presupposed* by our explanans (the phenomenological objectivity of our experiences) is a *virtue* of the proposed account of self-knowledge and not a defect in it. No explanatory circularity is involved.

Yet, although we may now be able to see how a pre-reflective form of self-consciousness can exist at the first-order level given that we experience the world as objective, the question of how we can experience the world as objective in the first place still remains unanswered. By putting the burden of explanation on the phenomenological objectivity of our experiences, we may not so much have explained how being implicitly self-conscious at the first-order level is possible as shifted our

question of how an implicit form of *self-consciousness* at the first-order level is possible to the equally puzzling question of how experiencing the world as *objective* is possible. In other words, how, one might still wonder, is such phenomenological objectivity, or even a conception of objectivity, possible in the first place?

To pose this question is it seems again to misunderstand the puzzle in question and the nature of the solution here on offer. The puzzle being solved by appeal to the phenomenological objectivity of our experiences is not a *developmental* puzzle but a puzzle about how given that we *do* have a grasp of objectivity and of subjectivity, and given that we *do* experience the world as objective, an implicit form of self-consciousness can be seen to exist at the first-order level. It begins from a being who is fully rational, whose attention is turned to the world, who possesses all the first-order and second-order concepts required for specifying the contents of her experiences and for self-ascribing these experiences, and who then moves directly from, say, a conscious experience as of a book on a table, to a judgement about her mental states of the form 'I believe that there is a book on the table' or 'It appears to me that there is a book on the table'. What is being explained in other words by reference to the phenomenological objectivity of our experiences is not how a pre-reflective form of self-consciousness might have *come about* but how its existence can be seen to fit in with how things currently *are*. The fact that this explanation does not tell us anything about how things might have *come to be* the way they are (ie. how we might have come to experience the world as objective) is thus not a defect in it. No explanatory regress is being generated.

In sum, what has been done in this chapter in appealing to the objectivity of our experiences and in drawing attention to the fact that this involves drawing a distinction between the objective and the subjective, is just to bring out how being implicitly self-conscious can indeed be seen to be present already in the way in which we *do* experience the world. The question of how we might have gone from being infants *not* experiencing the world as objective and *not* being self-conscious to being both self-conscious and experiencing the world as objective is another question altogether. This latter question, or indeed that of how we might have gone from possessing neither first nor second-order concepts to possessing both a sophisticated conception of the world as objective and of ourselves as having a perspective upon it, is a question which would require a *developmental* answer, or a suggestion as to how drawing a distinction between the objective and the subjective in experiencing the

world might be innate, combined perhaps with an account of how our visual systems parse the visual array, etc. No such answer is needed here however, as our question is, to repeat, not a developmental one but one about how an implicit form of self-consciousness (required for the possibility of fully reflective introspective self-knowledge) can be seen to fit in with how things actually are when we experience the world, and when, in so doing, we come to self-ascribe a perspective upon this world.

With this clearer idea in mind of the precise role that experiencing the world as objective and thereby being implicitly self-conscious is supposed to play in our solution to the puzzle about mental self-ascription, we can now go back to address a question left to us from the last chapter – that of the exact sense in which this so called 'implicit' form of self-consciousness that comes with experiencing the world as objective can truly be said to be *implicit*, and so to be something which can provide (as required) the *grounding* for our introspective second-order judgements, without being already itself a form of introspective self-knowledge.

5.4 'Pre-reflective' self-consciousness and 'non-conceptual' self-consciousness

In the last chapter we considered and rejected most possible interpretations of the notion of 'implicit' or 'pre-reflective self-consciousness' that could be extracted from the already existing literature on self-knowledge (from 'higher-order thought' accounts of consciousness and from a number of traditional accounts of self-knowledge – ie. perceptual accounts and constitutive accounts)[1]. So, how else could this notion be understood? One particularly similar sounding notion to our Sartrian one of 'pre-reflective self-consciousness' it seems still remains. It can be found in recent writings on 'non-conceptual' or 'pre-conceptual' self-consciousness.

Should we thus perhaps take 'implicit' or 'pre-reflective' in the present context to mean 'non-conceptual' or 'pre-conceptual'? If we should, might this not however open the door to a host of further explanatory questions about how one can move directly from a state with a second-order content that is non-conceptual to a fully conceptual articulation of this content in a second-order judgement? On the other hand if we should *not* take being 'implicit' here to mean being 'non-conceptual', in what other sense could our first-order states be said to be states of 'implicit' self-

[1] See 4.1 above

consciousness? In particular, how could a form of self-consciousness be said to be both implicit and *conceptual*? What could it possibly be to be *implicitly* deploying a second-order concept when attending to the world?

In what follows, it will be argued that the reflective/pre-reflective self-consciousness distinction we need to draw in this thesis does not map clearly onto the conceptual/non-conceptual self-consciousness distinction, and that in fact the whole present discussion about reflective and pre-reflective self-consciousness should, in a way, be thought of as orthogonal to debates about conceptual and non-conceptual content. This will not be to deny that there might be such a thing as non-conceptual self-consciousness, or to suggest that postulating such a form of self-consciousness might not be required for solving a number of other problems. The sole purpose of the comparison here will be to establish and specify, through this comparison, at least one distinctive sense in which a certain class of our states *must* be states of 'implicit' self-awareness regardless of whether there may also be other senses in which certain states of ours or of other creatures can or should be thought of as states of a 'less-than-fully-fledged' form of self-consciousness.

So, the broad question before us is this: in exactly what sense should we think of the 'implicit' or 'pre-reflective' form of self-consciousness required for the purposes of solving our puzzle as being *implicit*? More narrowly, should we take being 'implicit' or 'pre-reflective' in the context of this thesis to mean being 'non-conceptual' or '*pre*-conceptual'? That is, should the 'pre-reflective' form of self-consciousness which it has been argued *must* be involved in having all conscious states which we are able to immediately self-ascribe, be thought of essentially as a form of *non-conceptual* self-consciousness?

As mentioned, this section will essentially argue that issues about the conceptual or non-conceptual character of the content of the form of self-awareness we have in experiencing the world as objective should be thought of as quite orthogonal to issues about its implicitness or explicitness. That is, the notion put forward in this thesis of 'pre-reflective' self-consciousness should *not*, it will be argued, be thought of as equivalent to that of 'non-conceptual' self-consciousness, at least given a certain understanding of 'non-conceptual self-consciousness', namely the one that follows:

On the view to be considered here for heuristic purposes,[1] an attitude towards the world can be said to have *conceptual* content, if, in order to be in that state, one must actually possess the concepts required to specify that state's content. The content of my conscious thought that there is a book on this table will thus, for example, count as *conceptual* since I could not it seems be consciously thinking to myself that there is a book on this table unless I possessed the concept of a book, that of a table, the demonstrative concept *this*, the concept of being *on* something, and so on. On the other hand, a state can, on the understanding to be appealed to here, be said to have *non-conceptual* content, if one need not (though one might) possess the concepts required to specify the content of this state in order to be in it. It is often claimed that in this sense the contents of perceptual states are always fully or partially non-conceptual. Someone could for instance on this view be counted as having a perceptual experience as of something red (ie. a perceived object could appear to them *as red*) without them possessing the concept of 'red'.[2]

Assuming that there can be such things as states with non-conceptual content in the above sense, one might then go further and suggest that in addition to there being states with non-conceptual *first*-order contents (ie. contents to specify which one would need to explicitly deploy first-order concepts), there can also be states with non-conceptual *second*-order contents. Such a suggestion is for instance made by Bermudez, who argues that we need to posit the existence of such states for a number of reasons, in particular it seems the following three.[3]

In brief, one reason is essentially developmental, arising from the general question of how one can possibly acquire a concept unless one is, prior to that, already able to detect the presence of instances of the property designated by that concept. Another motivation is supposed to be that of making it possible to provide a non-circular account of what possessing second-order concepts, or in particular what possessing the 'I' concept, *consists in*. And, the third reason has to do primarily with

---

[1] See for instance (Crane 1992); also (Bermudez 1998, Chapter 3.)

[2] To give ourselves a frame of reference, Peacocke (1992) amongst others seems to believe in non-conceptual contents in the sense outlined here whereas McDowell (1994) for instance does not.

[3] Bermudez's concerns are more generally focused on establishing the existence of states with non-conceptual first *person* contents, whether these be *first-order* (as in the case of a content expressible by, say, 'I am standing in front of a tree') or *second-order* (as when judging 'I believe that it is raining'). The arguments he provides, and psychological evidence he appeals to are thus not always explicitly put forward in support of the existence of specifically *psychological* first person contents, though I will of course be focusing primarily on these latter cases.

trying to make the best possible sense of some of the behaviour of animals and very young children who clearly do not possess any second-order concepts, yet whose behaviour in certain circumstances is such that the principle of inference to the best explanation seems to require attributing to them states with, for one thing, *content* (that is, it seems to require interpreting their behaviour as not purely reactive, but as based on their actually *representing* things to be one way or another or experiencing things *as* this or that), and secondly with *second-order* contents.[1]

Now, without going into any of these arguments in great detail – since it is not amongst our present aims to assess the case for the existence of states of non-conceptual self-consciousness but merely to compare the notion of 'non-conceptual self-consciousness' with that of 'pre-reflective self-consciousness' set out in this thesis – let us just briefly go over some of the empirical evidence appealed to, so as to have in mind some concrete examples of what states of 'non-conceptual self-consciousness' are supposed to be.

One source of empirical evidence for the existence of such states is supposed be found in experiments of so called 'joint visual attention', where inference to the best explanation of the observed behaviour of pre-linguistic children in interaction with their mothers, seems to require attributing to the infants in question states with (a) *contents* and (b) with contents which we would normally express by such sentences such as 'Mother wants me to look where she is looking' or 'I am looking where Mother is looking' or 'Mother will look where I am looking if I look back and forth from her to it', etc. Now, if this is indeed the right interpretation to give of these infants' behaviour in these experiments, we would in fact seem to have here paradigm cases of states with non-conceptual second-order contents in the sense defined above, since the infants are supposedly able to be in states with the above second-order contents without possessing any second-order concepts, and indeed without possessing any concepts at all.[2]

---

[1] Incidentally, in support of the claims made earlier on, beings who do not possess any second-order concepts do not seem to possess any first-order concepts either.

[2] See (Bermudez 1998, ch.9) for a more detailed discussion of these experiments and their potential support for the existence of states with non-conceptual second-order contents. As already mentioned however, I am side-stepping all these details here essentially because it is not my present aim to take sides on the issue of what these experiments can actually show, or indeed on the issue of the possibility or not of there being states with non-conceptual content, whether at the first or the second-order level. The primary aim of this section is to suggest that we *do not need* to take sides on the question of whether there can be states of non-conceptual self-consciousness, or indeed states with non-conceptual content at *any* level in order to make room for the existence of '*pre-reflective*' or *implicit* self-consciousness in the sense needed to solve our puzzle about mental self-ascription.

But now, given the above examples of states of *non-conceptual* self-consciousness, and given what has been said in this thesis about *pre-reflective* self-consciousness and the specific explanatory gap it is supposed to fill, it seems quite clear that the notions of 'non-conceptual self-consciousness' and 'pre-reflective self-consciousness' will fail, at least in some cases, to match up in their extensions. In particular, what is needed to solve *our* problem is a form of self-consciousness that constitutes an *intrinsic* part of what is involved in being in an essentially *first-order* conscious state. It is not at all clear however that the states in the above examples of joint visual attention can be thought of as first-order states. They are states about one's own mental states (ie. *second-order* states). To this extent, the notion of 'pre-reflective' self-consciousness needed to solve our puzzle is much closer to the notion of self-consciousness sketched by Sartre in his talk of the 'pre-reflective cogito' (which he stresses is not a separate state *about* a first-order conscious state, but part of the very mode of being of the first-order conscious state itself)[1] than to Bermudez's notion of 'non-conceptual self-consciousness'. In the above examples at least, the states of supposed 'non-conceptual self-consciousness' in question are not first-order states of implicit self-awareness but clearly *second-order* states (ie. states with contents to specify which one would need to explicitly deploy *second-order* concepts). If these are therefore supposed to be paradigm examples of states of 'non-conceptual self-consciousness', then positing the possibility of a non-conceptual form of self-consciousness (ie. a form of self-consciousness which does not require possessing or deploying second-order concepts) cannot alone get us any closer to understanding how self-ascribing a conscious state directly on the basis of having it (ie. on the basis of attending to the *world*) is possible. The examples of non-conceptual self-consciousness taken from experiments of joint visual attention seem to be essentially cases of what we might call *explicit* yet *not conceptual* thoughts about one's own mental states, insofar as what the infants' attention in these examples is occupied with are facts about *themselves looking* (or their mother looking) rather than facts about how the *world* is,[2] thereby suggesting that there is a possible distinction to be drawn here between the idea of a form of self-consciousness being

[1] See (Sartre 1969, Introduction)

[2] By 'attention' here I do not of course mean perceptual attention or perceptual focus, which, either way, is going to be on some aspect of the world, if on anything.

implicit and it's being *non-conceptual*. It's implicitness has to do with what one is or is not *attending to*, whereas it's non-conceptual character has to do with what concepts one need or need not possess in order to be in that state.

This is of course not to say that there could not be states of self-awareness which are both implicit and non-conceptual. An illustration of such states of implicit non-conceptual self-consciousness can in fact be found in another source of evidence appealed to by Bermudez for the existence of states with non-conceptual first-person contents, namely in Gibson's theory of ecological optics. Based on various experiments performed on non-linguistic animals and infants, Gibson suggests that outward-directed perceptual experience itself, even in the case of beings with no conceptual abilities at all, is a direct source of information not only about the world, but also about oneself: of one's location, of one's own movement and position, of various possibilities regarding what one can or cannot do, etc.[1] To take an example, this is supposed to be shown by the fact that in so called 'moving room' experiments (where infants are put in a room the walls of which are made to move) the infants involved, by looking only ahead of themselves, tend to take the changes in their visual field to signal a change in their *own* position and movement. This is supposedly revealed by their tendency to compensate for these changes by moving their bodies in the opposite direction and eventually falling over.

Now, although it is not clear that these kinds of considerations alone can actually show that the relevant animals or infants are in anything like contentful states (rather than manifesting a perhaps surprising but purely practical ability to attune themselves to their environment), and although it is not clear either that the self-specifying information here is in any straightforward way information about their *mental* states, it certainly seems that being capable of reacting in this way to perceived changes in their visual field does display *some* kind of sensitivity to the distinction between the perceived world and their subjective point of view upon it. So, perhaps, these cases of infants co-perceiving the world and properties of themselves as well as making some subjective/objective distinction through attending primarily to the environment, could be taken to be examples of them being in states of *implicit, non-conceptual* self-consciousness – implicit because their attention is, if on anything explicitly, on the environment (rather than on their perceptual experience), and *non-*

[1] See (Gibson J. J 1979 as well as Gibson, E. J. 1969) and (Bermudez 1998, Ch. 5)

conceptual because of course they do not possess the concepts needed to specify or represent any of this conceptually (whether the world or the fact of their looking at it at a certain angle).

Nevertheless, although such an implicit form of self-consciousness might indeed be involved in having these first-order perceptual states (first-order states which are themselves non-conceptual), and although the form of self-consciousness these infants would seem to have is certainly going to have to be a form of *non-conceptual* self-consciousness (since they do not possess any second-order concepts), it is not clear that in our own case (ie. in adults), given what has been suggested about the inter-dependence of first and second order concepts,[1] the implicit form of self-consciousness that comes with having first-order states with *conceptual* content (ie. of a kind that we at least sometimes have) can come out as being a form of *non-conceptual* self-consciousness. Being pre-reflectively self-conscious in the way that comes with being in a first-order state with *conceptual* content will of course not involve *explicitly* deploying any second-order concepts (since in having a first-order state one is attending only to the *world*), but it certainly will require *possessing* such concepts, since, if the suggestion about the interdependence of first and second-order concepts is correct, possessing some second-order concepts will be *presupposed* by one's possession of the concepts actually deployed at the first-order level.

To try to make this latter point clearer, and to avoid getting ourselves tangled up in ambiguous terminology, let us remind ourselves of the two main factors at work here: (1) a state is supposed to have conceptual/non-conceptual content if being in a state with that content requires/does not require possessing the concepts needed for specifying that content, and (2), following the suggestion made in section 5.2 above, one cannot possess sophisticated first-order concepts of the kind that we have, without also possessing some second-order concepts. Now, if both of these claims are true, it would seem that one cannot be in a first-order state with *conceptual* content without possessing any second-order concepts, and can therefore not be 'pre-reflectively' self-conscious in the way that one actually *is* when in a first-order state with *conceptual* content, without possessing some second-order concepts. So, in a somewhat indirect sense, the pre-reflective form of self-consciousness involved in having first-order conscious states with conceptual content (ie. of the kind that we at least sometimes

[1] See pp. 110-112 above

have) requires possessing second-order concepts, and thus, on the above definition of conceptual and non-conceptual content, this form of self-consciousness would come out as being *conceptual*, although being self-conscious in this way clearly does not involve *explicitly* deploying any second-order concepts. Perhaps the best way of putting it then would be to say, if we can make sense of this idea at all, that it involves *implicitly* deploying these concepts. It is a form of implicit, yet conceptual, self-consciousness.

But, does it actually make sense to talk of *implicitly* deploying a concept? It is not clear to me that it does not. There are in fact other cases in which it would seem appropriate to say that a concept is being deployed implicitly. Let us digress for a moment and consider a discussion in the philosophy of time.

Consider the case of someone saying 'I am writing an essay at the time of this utterance'. This is something along the lines of what reductionists about dynamic temporal facts (ie. facts about pastness, presentness and futurity) to static temporal facts (ie. facts about simultaneity and being earlier or later than some event) generally want to say that we actually mean when we say such things as 'I am *now* writing an essay'.[1] A good objection to such a view is however to say that, insofar as this analysis works, it is not in fact a reductive analysis at all, since what is meant by '*this* utterance' is precisely 'the utterance which I am making *now*'. It is a distinctive feature of demonstrative reference with the word 'this' that it presupposes reference to the temporal presence of that which is being picked out. Thinking of an event *as* '*this* utterance' is to be thinking of it as the utterance which is occurring *now*. Thus, assuming that this analysis is correct, it does not seem that one could fully grasp the content of what one is saying in saying 'I am writing an essay at the time of *this* utterance' without having some grasp of the concept of something's occurring *now*. Consequently, in a sense, saying something like 'I am writing an essay at the time of this utterance' could be said to involve *implicitly* deploying the concept of temporal presence 'now', or to involve making *implicit* reference to temporal presence, in a way that on the other hand saying something like 'I am writing an essay at time t' or 'I am writing an essay at the time of my utterance of 'I am writing an essay'' would not.

[1] For various reductionist and non reductionist views of temporal facts, see LePoidevin and MacBeath (eds.) (1993), section 1.

Incidentally, referring demonstratively seems to presuppose not just implicit reference to dynamic temporal facts, but also to ourselves and to our mental states, such as for instance when pointing at something in our immediate vicinity and saying '*that* object' meaning essentially 'the object *I am now pointing* at', or 'the object in my *visual field*' or 'the object I am now *looking* at'. Similarly, in non-perceptual cases, when one is for instance just reflecting on something and suddenly comes to think '*that* is interesting', what one seems to mean by 'that' is essentially 'the proposition that just *occurred to me*' or 'the thing I just *thought* of'. These considerations may I believe be of some importance when thinking about our ability to self-ascribe thoughts that are not about the world directly before us, but, say, about abstract ideas or propositions.

To return to our discussion of non-conceptual self-consciousness, the point to take home is that what is essential to a first-order state's being a state of *implicit* or *pre-reflective* self-consciousness in the sense needed to solve our puzzle in this thesis, is that this self-awareness be involved in the very having of the first-order conscious state itself, and not that being in this state not require possessing second-order concepts. In fact the pre-reflective form of self-consciousness involved in having first-order states with *conceptual* content (ie. some of the very states we are able to self-ascribe directly on the basis of having them – eg. our occurrent empirical thoughts) would even seem to *require* possessing second-order concepts, since, as we have seen, being in a world-directed state with conceptual content not only involves being pre-reflectively self-conscious, but also, or so I have suggested, presupposes possessing second-order concepts. Thus, if there is to be such a thing as *non-conceptual* pre-reflective self-consciousness, it will have to be (as perhaps is the case of the animals and infants in the Gibsonian experiments) essentially associated with being in states with *non-conceptual* first-order contents.[1] In any case, whatever side one chooses to take on the issue of the existence or not of states with non-conceptual contents (whether at the first or second-order level), no threat is going to be posed to the present solution to our problem about mental self-ascription. The side one chooses to take on this issue may well change what one can say about animals and very young children, but it will not change anything essential, in most cases, about what the situation is for us.

[1] This again fits quite well with the idea that the kind of grasp one has of oneself and of one's subjectivity will correspond to the kind of grasp one has of the world and of its objectivity. See pp. 112-113 above

Of course, one could mean something very different by 'non-conceptual' than that which I have been assuming it to mean here. In fact, it might be claimed that 'non-conceptual' just means 'implicit' or 'not fully spelled out' or 'not currently attended to' thereby affecting whether we should label the form of self-consciousness I have been calling 'pre-reflective', 'non-conceptual'. This would however be a purely terminological issue. If what one means by 'non-conceptual' encompasses all that I have been meaning here by 'implicit' or 'pre-reflective', then I am happy to call this form of self-consciousness 'non-conceptual'. It is far from clear however that this is what is generally meant by 'non-conceptual'. More relevantly, the main purpose of the above comparison between the above two notions of self-consciousness ('pre-reflective' and 'non-conceptual') was simply to allow us to articulate more clearly what 'pre-reflective' self-consciousness is supposed to be in the context of solving our puzzle about mental self-ascription, regardless of what might or might not be meant, in other contexts, by 'non-conceptual self-consciousness'.

To sum up, what we have done in this chapter is put forward and clarify a concrete solution to the puzzle of how our introspective self-ascriptive judgements can be based directly on looking out at the world, and thereby ultimately provided a solution to our original problem of how immediate, authoritative, immune to non-cognitive error knowledge of our own conscious mental states is possible. Or, we have at least done so for the case of world-directed cognitive states.

The focus of our solution has been on the phenomenological objectivity of our experiences, thoughts, etc. as the one aspect of our first-order conscious states in virtue of which they can be states of 'pre-reflective self-consciousness' and thereby constitute immediate rational grounds for self-ascribing them. This still leaves a fundamental question unanswered – that of how we might be able to self-ascribe, in a special authoritative way, states of ours other than beliefs, experiences, thoughts, etc. as of a mind-independent world. More precisely, the problem left before us is this: if what has been argued in this thesis so far is correct, all mental sates which we are able to authoritatively self-ascribe, we must be able to self-ascribe *directly on the basis of having them*, that is, directly on the basis of attending to their objects out in the world. Yet, it is not clear (a) that in having all conscious states, in particular certain emotions and desires, their objects always present themselves to us as objective, nor (b) that/how our experiencing the world as objective, even when doing so *is* involved in our having particular emotions and desires, can involve our being implicitly aware of

ourselves as having distinctively *emotional* attitudes towards the world (eg. fear, hope, etc.) as opposed to just a cognitive perspective upon it. In sum, it is far from clear how the suggestions made so far about how a state can be a state of pre-reflective self-consciousness can be extended to cover the case of all conscious states of which we are able to have immediate authoritative knowledge.

But, perhaps these suggestions do not *need* to be so extended. Perhaps, that is, we are not actually able to know our own emotions and desires in the special authoritative way in which we are able to know our own cognitive states, and so do not need to show how attending to the world when having an emotion or desire might involve being pre-reflectively self-conscious. Or do we? In order to answer this question, a step back will need to be taken to re-consider the puzzle about mental self-ascription to which pre-reflective self-consciousness was supposed to be an answer, and to focus this time on how a parallel puzzle might arise, if at all, for the case of states other than thoughts and experiences as of a mind-independent world.

Chapter 6: Knowing our own emotions

On the face of it, the process of self-ascribing our own emotions is quite different from that of self-ascribing our own cognitive states. Self-ascribing our own emotions and desires tends in particular to be a far lengthier, more difficult, and more prone to error process than that of self-ascribing most of our own thoughts, beliefs and perceptual experiences. This would initially seem to suggest that a very different account of self-knowledge should be adopted for the case of emotions and desires from that adopted for the case of cognitive states. Take for example the process of self-ascribing a feeling of guilt about feeling envy. Such a process is likely to be both lengthy and highly susceptible to error very much *un*like a process of self-ascribing, say, a simple belief (eg. that it is raining) or an occurrent perceptual experience (eg. as of a table). Should this not be taken to suggest that although we are able to know our own cognitive states in a way that is immediate, authoritative, and immune to non-cognitive error, we are perhaps *not* able to know our own emotions in any such special way? Should we not in fact adopt something more like a Rylean account of self-knowledge for the case of emotions and desires than we have adopted for the case of cognitive states – ie. an account according to which we come to know our own emotions and desires in much the same way as we come to know the emotions of others, namely by inference from observation (or from direct observation) of our behavioural patterns?

The view that first-person authority begins and ends with our knowledge of our own consciously manifested cognitive states and perceptual experiences, to the exclusion of motivational states like emotions and desires, is initially quite compelling

in that it seems to fit well with the innumerable situations where we do not appear to know what we want, where we are not sure how we feel or *are* sure but are in fact mistaken, or where we get things right only at cost of much effort and lengthy interpretative reflection. It could in fact quite plausibly be argued that the whole problem about introspective self-knowledge (and resulting puzzle about mental self-ascription) does not arise at all for the case of emotions and desires. Contrary to this however, it will be argued here that a parallel puzzle about mental self-ascription *does* arise for the case of emotions and desires, and that a parallel solution therefore to it must be found – ie. a solution that can show how attending to the objects of our emotions and desires involves, like attending to the objects of our cognitive states, being implicitly aware of ourselves as having them. So, how might such a puzzle arise?

6.1 The puzzle

At first glance, as mentioned above, there is much data to suggest that unlike our own cognitive states, we do *not* have immediate authoritative access to our own emotions and desires. Consider for instance what you want to do with your life, or whether you want to do a particular thing with it such as have children, change job, etc. Alternatively, ask yourself whether you resent, say, your family for preventing you from following a certain path in your career, or whether it is instead yourself that you are angry at for not having pursued it. Even more simply, ask yourself whether you are happy, or whether you want to go to the countryside for the weekend, or indeed whether you like the colour of the wall in front of you. You may in all these cases be initially quite unsure what to say. In some of them, you may even remain unsure after suitable consideration and perhaps even get what you feel wrong in the end, as an analyst, or someone who knows you well, might make you realise later. In other cases you may be wrong even when initially certain that you do (or do not, as the case may be) feel a certain way about something. For instance, you may be convinced that you have no feelings at all for someone, yet find yourself trembling nervously upon suddenly seeing them walk into the room. Or, more interestingly, you may come to this realisation that you still love/hate/are angry at this person little by little, upon observing them, thinking about them, recalling past situations involving them, listening to what they are saying, etc.

These examples are of course different in many ways. They all seem to suggest however that in one way or another the process by which we come to know our own emotions (even in cases of emotions which clearly are not unconscious or repressed) tends to be one fraught with the possibility of error, and even when carried through successfully, often a laborious task. None of this seems true of most cases of *cognitive* self-ascriptions. If, for instance, instead of being asked how you feel about someone you are asked whether you are now *thinking* about them, you will probably be immediately able to say that you are or are not thinking about them, without much hesitation. And, except in rare circumstances (eg. of irrationality, insincerity, or conceptual incompetence), you are not likely to be mistaken. The question therefore is why are things different in cases of self-ascription of emotions and desires? The obvious answer is to say that things are different because we do not actually come to know our own emotions and desires directly on the basis of having them as we do many of our cognitive states, but only on the basis of a long and interpretative inferential process. That is, the obvious suggestion is to say that the process by which we reach judgements about our own emotions is not a directly introspective one like that of self-ascribing our own cognitive states, but a complex inferential and interpretative one, based essentially on observing our own behavioural and nervous system reactions, noting recurring patterns of (non-emotional) thought, noting what we are perceptually attending to, and so on. Some of this data may of course be acquired through direct introspection, particularly the data about what thoughts we are having, what we are perceptually attending to, what we are imagining, and so on. But, this in itself does not make the knowledge we thereby acquire of our own *emotions* non-inferential or introspectively based in the sense described in the first part of the thesis. In fact regardless of how the information about ourselves is acquired, it remains information about *ourselves* and not about the *world* (only the latter would give rise to our puzzle), and, once acquired, can only be used in the way a third person would use it, essentially in a process of *inference* to derive what emotions we have.

So, why should having emotions and desires actually need to be shown to involve our being implicitly aware of ourselves as having them? Why, in fact, should we be taken to have any kind of special access to our own emotions at all?

It certainly seems true that some of our emotions, in particular our repressed or unconscious ones (in the Freudian sense), are known by us only in the above way, that is, not through any direct form of introspection, but essentially through inference from

explicit data regarding ourselves. I might for instance have a repressed desire to hurt someone (a desire which might be repressed, say, because of some other emotion I also have – eg. a feeling of guilt about having such a desire), which therefore fails to manifest itself in my conscious mind (eg. in what I consciously believe about the person or in how I intentionally relate to them), but which I may nonetheless come to realise that I have through noting various involuntary behavioural and nervous system reactions on my part in this person's presence, or by noting recurrent thoughts or dreams I might have involving them. Taking *all* emotional self-knowledge however to be inferential in this way ultimately fails, we will see, to do justice (a) to certain clear distinctions that exist between different cases of self-ascription of emotions (eg. between self-ascriptions of conscious emotions and self-ascriptions of unconscious emotions), and (b) to certain important *similarities* that exist between some cases of self-ascription of emotions and a number of cases of introspectively (ie. *not* inferentially) based self-ascriptions of cognitive states and perceptual experiences.

To begin with, it is not just some of our own emotions and desires (ie. our unconscious or repressed emotions and desires) that we are able to know inferentially, but also a wide range of our own *beliefs* – ie. our repressed beliefs such as in cases of denial. The latter fact, ie. that some of our beliefs are known only inferentially does not however, as discussed earlier in the thesis, preclude other beliefs of ours from being accessible to us through direct introspection (or at least through direct introspection of their conscious manifestations in judgement). Thus, similarly, there is no reason yet to believe that all our emotions are known only inferentially. Some of our emotions and desires, like some of our beliefs, may well be known in a special first-person authoritative manner. Next, it so happens that many conscious (in the dispositional sense) emotions do seem to function quite differently from *un*conscious (ie. repressed) emotions. They seem in particular capable of manifesting themselves in episodes of consciousness (in our occurrent thoughts, experiences, episodes of visualisation, etc.), and crucially, they seem knowable (like many of our cognitive states) *without* the need for any inference from data explicitly regarding ourselves.

Take for instance the desire you might now have for chocolate cake, or your love for someone, or your embarrassment about something you did last night, or your mild annoyance at the person in front of you partially blocking your view at the cinema. In all these cases, if asked whether you have the relevant emotion, you are likely to turn your attention to the objects or events themselves rather than to yourself

and to your behaviour, arousal symptoms, patterns of non-emotional thought, etc. Suppose in fact that you are now presented with the chocolate cake and asked whether you would like some. You will probably not turn your attention to how you are behaving (you may be sitting perfectly still), or to whether you are having any specific arousal symptoms (you may not be having any, or at least not any specific enough to identify yourself as desiring chocolate cake rather than as having some other emotion), nor indeed will you introspect any relevant non-emotional thoughts you might be having (you may well not be having any explicit thoughts about the cake's being good or worth having, etc. let alone be actually consulting these thoughts). Instead, what you *will* it seems most likely do upon being offered chocolate cake is consider the *cake* (not the fact that your are considering it), turn your attention and gaze towards *it* (not to the fact that you are looking at it or visualising it in imagination), and so on. Similar things could also be said of the other examples. In response for instance to the question of whether you are embarrassed about something you did the night before, you will probably run through the events of the previous evening in your head, focusing on particular moments of it, rather than turn your attention to your present behaviour, thought patterns, or bodily sensations. Of course in many cases of self-ascription of emotion the latter kind of evidence is consulted *as well*. The point however remains that often it is *not* consulted and that in some cases it *could* not be consulted not being actually available, so it obviously *need* not be consulted in all cases. What we do need to attend to in these cases however is it seems the *objects* of our emotions (whether they be objects, states of affairs, events, abstract possibilities, etc.). Our knowledge of our own emotions in these cases therefore has to be explained other than by way of an inferential account according to which our knowledge of our own emotions is derived essentially from explicit data regarding ourselves.

It now seems that there is clearly an epistemological difference between different cases of self-ascription of emotion – between instances of self-ascription of unconscious emotions and instances of self-ascription of consciously manifested (or at least consciously manifestable) emotions – and that this difference lies specifically in what kind of evidence is consulted in each case. Moreover, the evidence in play in coming to know the latter set of emotions, seems to be very much akin to that in play in coming to know our own conscious (ie. non-repressed) *cognitive* states and perceptual experiences. In both cases, in being asked whether we have a particular

state (emotional or cognitive), we seem to focus our attention on the world rather than on ourselves and on our behaviour, arousal symptoms, or other mental and physical properties. Our judgements about our own conscious emotions thus seem to be in this respect reached on a similar basis to those about our conscious beliefs, ie. on the basis of *having* first-order occurrent states (in which our emotions/beliefs are manifested), rather than on the basis of thinking about the fact that we are having them. This is clearly something which no third person is in a position to do, and so suggests that in the end, as with the case of beliefs, we *do* have some kind of privileged access to a wide range of our own emotions. No one other than ourselves can effect a direct psychological transition between *having* a mental state of ours and judging that we have it, since only we ourselves have our states.[1]

With this similarity though between our knowledge of certain of our emotions and our knowledge of some of our cognitive states comes also a similar puzzle: how can *having* a first-order consciously manifested emotion itself already reveal to us the fact that we have it? Or, more specifically, how can attending to the *object* of an emotion of ours directly reveal to us the fact that we have an *emotional* attitude towards it rather than just a cognitive perspective upon it? The same puzzle that arose for the case of cognitive states in other words clearly arises also for the case of emotions and desires – or at least a very similar puzzle which will require a similar solution.

It is true, it must be said, that in self-ascribing emotions, we do not (as we do in self-ascribing experiences and beliefs) actually consider *whether* the world is a certain way. That is, we do not make self-ascriptive judgements about, say, whether we dislike rain by looking at whether it is actually raining. We do however *attend* to the objects of our emotions in order to determine whether we have these emotions, and so do move directly from the first-order emotional level to the second-order self-ascriptive level, that is from our emotional perspective on the world to self-ascribing an emotion towards it. The puzzle about self-knowledge thus arises nonetheless, and still requires the same kind of solution. In fact, additionally, just as when self-ascribing a cognitive state, we turn our attention not only to the world in coming to

[1] There is of course a sense in which the *fact* of one's having a certain emotion could constitute a direct 'reason' (relative to some normative system) for a third person to move on to judge that one has it. In the sense of transition appealed to in this thesis though, as should be clear from what was said in previous chapters about reasons (see in particular 3.1 above), the transition in question is an essentially *psychological* one, holding between the mental occurrences themselves, and not a normative relation between the *fact* that the one state is occurring and someone's believing that it is.

self-ascribe an emotion, but doing so seems generally to *make sense* to us from within our own conscious point of view. We do not just 'find ourselves' self-ascribing emotions on particular occasions. Rather, when self-ascribing an emotion based on looking out at the world, doing so seems *appropriate* to us from within our own outward-looking perspective. Our attention to the world does not just act as a *trigger* for our self-ascriptions. Rather, we seem to turn to the world for *evidence* or *confirmation* that we have a particular emotion.

In sum, despite the differences that clearly exist between our knowledge of our own cognitive states and our knowledge of our own emotions, a crucial point of similarity between the two is also striking. When self-ascribing a wide range of our own emotions, as when self-ascribing a wide range of our own cognitive states, we tend to turn our attention not to ourselves (eg. to our behavioural patterns) or to our mental states (ie. to psychological facts about ourselves) for evidence or confirmation, but primarily to the *world*, that is, to the *objects* of our attitudes, or to *that towards which* we have these mental states. The puzzle about mental self-ascription arrived at in the preceding chapters arises in other words both for the case of our knowledge of a certain class of our own cognitive states (beliefs, thoughts, perceptual experiences, etc.) and for the case of our knowledge of many of our own emotions (eg. desires, fears, hopes, etc.). An account of how implicit reference to ourselves as having a particular attitude towards the world can already be present in the way the world strikes us phenomenologically does therefore turn out to be needed for both cases. In light of the differences however between emotional states and cognitive states, and between our abilities to *know* each of them, a slightly different account may in the end turn out to be needed for the case of the latter from that adopted for the case of the former. Broadly though, the correct account of self-knowledge for the case of emotions will it seems have to be, as it was for the case of cognitive states, an account of how our having these attitudes towards the world can be already implicit in the way the world (or the objects of these attitudes) strikes us in our having these attitudes. A difficult question remains though about how this might be done.

6.2 Knowing our own emotions and experiencing the world as objective

It was suggested, recall, that our world-directed judgements and perceptual experiences could be understood to be states of implicit self-awareness by appeal to

the distinctive self/world dual way in which the world presents itself to us (or is represented as being) in our having these experiences, thoughts, etc. Put differently, the starting point of our solution to the puzzle about mental self-ascription for the case of cognitive states was simple. It was a plausible assumption about the phenomenology of world-directed thought and experience, that is, the assumption that, phenomenologically at least, we experience and think about the world as *objective*, or, put differently, that our first-order thoughts, experiences, episodes of imagination, etc. are *as of* (even if not necessarily *of*, or *believed* to be of) objective tables and chairs, etc. that is, in effect, as of tables and chairs upon which we have a point of view. Attempting to account for our knowledge of our own *emotions* and *desires* by appeal to the idea of our representing the world as objective comes however with an added challenge. Two problems in particular immediately present themselves.

First, it was argued that if a first-order mental state is to constitute a direct ground for authoritative knowledge of a given mental state (be it a cognitive state or an emotion) this first-order episode must either be the mental state to be self-ascribed itself or an episodic conscious expression of it. Yet, it is far from clear how episodes of representing the world as objective (or of it's directly presenting itself to us as objective) can possibly be thought of as episodes of *emotion*, or even as episodic *expressions* of emotion (if one wants to think of emotions as primarily dispositional states). Second, even if one grants that in some cases emotions *do* manifest themselves (in part at least – manifestations of emotions may be complex) in the world's striking us as being objectively a certain way (eg. paranoia may manifest itself in our representing someone as actually out to get us), a further problem arises, namely that of how representing the world in this way can possibly constitute, from within our own outward looking point of view, a direct source of knowledge of the fact that we are not just representing the world or certain aspects of it as being a certain way (eg. representing someone as out to get us) but have a specific underlying *emotion* (eg. paranoia) towards it.

Let us consider these two problems in turn, starting with a more precise look at the first.

(1) As already mentioned, authoritative self-knowledge is supposed to be knowledge based directly on the mental states thereby known, or in the case of dispositional states, knowledge based on episodic manifestations of these states in

consciousness. However, it is not at all clear that episodes of representing the world merely as being a certain way objectively can be thought of either as actual episodes of emotion or as episodic world-directed *expressions* of emotion (in the way that, say, certain acts of judgement might be seen as conscious expressions of belief). If instances of representing the world (or of its directly presenting itself to us in perception) as being objectively one way or another are thus to be shown to be the first-order episodes on the basis of which direct authoritative knowledge of our own emotions and desires is possible, these episodes will have to be shown to be in some respect conscious expressions of emotions and desires. There are however a number of reasons for thinking that they cannot be. Consider for instance the simple case of desire.

It is often assumed in the philosophy of mind that beliefs and desires are distinct types of attitudes but potentially towards the same objects (propositions or states of affairs) considered in the exact same way. That is, apart from the (not uncontroversial) assumption that all desires are in fact directed at propositions or states of affairs, it is assumed that with respect to any proposition p, or with respect to any particular state of affairs, one can have either the attitude of belief or that of desire. No difference in other words is taken to exist at the level of content but only at the level of attitude. Upon closer examination however, there does seem to be an important difference between what one desires and what one believes, or between how the objects of one's beliefs and desires appear in one's mind when one supposedly 'desires that p' and when one supposedly 'believes that p'.

Suppose for instance that p is the (possible or actual) 'state of affairs' of there being world peace. When believing that p, what one believes, and so what one consciously represents when the question arises, is, it seems, that there *is* world peace. That is, one represents world peace as an objective state of affairs, as something that actually obtains mind-independently. On the other hand, when desiring supposedly the same thing, ie. 'that p', that is, when having a desire supposedly towards the same proposition, fact, or state of affairs, what one seems to desire is not actually that there *is* world peace (ie. that anything is objectively the case), but that there *be* world peace. The idea of there being world peace seems to figure in one's mind very differently according to whether one has the belief or the desire. In the case of desire, when one's attention is focused on the idea of world peace (and so this desire – assuming it is not a repressed desire – comes to manifest itself in consciousness), one's attention is it

seems on world peace considered as a possible way the world could be (or perhaps as a way it is *to be made to be*) but not on world peace as a way the world actually is, ie. as a fact, or as a mind-independent state of affairs upon which one has a point of view. If so however, that is, if the objects of our desires do not feature in our attention as objective states of affairs, then episodes of representing the world as being objectively one way or another (such as when judging 'there is world peace' or 'there is food in the refrigerator') cannot it seems capture our world-directed perspective in desiring these things and so cannot be said to be direct expressions of these desires at the first-order level, nor ultimately therefore to be the direct non-inferential grounds on which authoritative knowledge of these desires could be based. If a mental episode is to constitute an expression or manifestation of a particular attitude at the conscious world-directed level, and so to be the possible ground for direct authoritative knowledge of this attitude, the episode must it seems at the very least capture how the object of one's attitude actually presents itself to one, from within one's conscious perspective in having this attitude and attending to its object. Instances of representing the world as objectively one way or another do not seem to be able to do this for the case of most desires.

One might go even further and say that representing the world as being a certain way objectively is not only unable to capture the way the world is apprehended from the point of view of desire, but that it actually goes *contrary* to some of what is directly implied by the way in which the objects of our desires figure in our minds from this perspective. To start with, there is already something distinctively inconsistent-sounding about saying such things as that one desires that there is world peace, or that one desires that it actually is raining. Not only is putting things this way ungrammatical, but it seems for good reason. As mentioned above, when one has the idea of world peace in mind in desire, one has it in mind that world peace be the case, not that it is the case. And, having in mind that something be the case seems to imply that it is not, at least from one's own epistemic standpoint, the case. There being world peace is represented merely as a possible way the world could be, and so implicitly not as the way things actually are. Far therefore from being a potential direct expression of a desire that the world be one way or another, having in mind that the world is objectively some way seems to altogether contradict the way things are represented in desire. If anything, representing the world as being

objectively a certain way suggests that one probably does not desire that it be that way (or no longer desires this, and is now just happy that it is so).

It should be noted however that this is not to say that there is nothing more to the difference between the attitudes of belief and desire than a difference in their contents or a difference in how their respective objects are apprehended or represented when attending to these objects. It is important however to realize that the difference between beliefs and desires does not hold only at the level of attitudes (whatever this difference may amount to) but also at the level of world-directed content, and indeed in more ways than one. Martin for instance goes further and argues that a number of desires are fundamentally desires towards *events*, and cannot be reduced to propositional desires, or to desires towards states of affairs without thereby rendering entirely unintelligible certain important aspects of the role of these desires in motivation, as well as certain crucial facts about the circumstances in which these desires will tend to cease or persist[1]. For example, if one has the desire 'to drink a beer', and this desire is reinterpreted as the desire that some particular state of affairs obtain, say, the state of affairs of one now drinking a beer, one would end up with no way of explaining why once now (ie. the present moment) has passed, and one has not yet had a beer, the desire should persist. Leaving aside though for the moment why we should think of such desires as fundamentally desires towards events rather than towards states of affairs, there is a feature these desires seem to share with the desires considered earlier. In having the desire 'to drink a beer', if this desire is taken at face value as a desire for just that, drinking a beer will it seems figure in one's mind as a general type of event, perhaps additionally as one to be actualised, but clearly not as any particular objective event actually occurring, whether now or at some specific future time. Thus, again, the objects of one's desires (be they events, states of affairs, or anything else) do not seem to figure in one's mind, from the perspective of desire, as objectively obtaining. This of course presents a serious problem (more than one in fact) for any attempt to extend the preceding account of self-consciousness to the case of desire.

To add to these problems, there being the above differences between the way things are represented in belief and in desire (as objective in belief, and specifically *not* so in desire) has much data to support it. It seems to fit well for instance with, and

[1] See Martin (1999)

may indeed help *explain*, a number of claims often made about beliefs and desires in the literature on 'direction of fit',[1] claims that would be difficult to explain on the assumption that beliefs and desires are attitudes towards the exact same things considered in the exact same way.

For example, one claim often made is that there is something wrong with or irrational about believing, say, that p, when the world presents itself to one in perception as *not*-p, whereas there is no cognitive failing or irrationality involved in *desiring* that p when in the presence of similar evidence, or when actually simultaneously representing that *not*-p. In light of the earlier distinctions, this difference might now be explained as follows. In both perception and consciously expressed belief (but not in desire) the world presents itself to one, or is represented, as being objectively one way or another. Perceiving that *not*-p can thus be seen to constitute an immediate reason for taking *not*-p to be objectively the case. If one were therefore to believe that *p* (ie. to take p to be the case) *despite* the perceptual evidence one has to the contrary, one could easily be seen as going against (or failing to take account of) one's *reasons* or *evidence*, and so as being *irrational* (assuming one has no reason to mistrust one's senses) or as subject to some other cognitive failing – eg. division of the mind, conceptual incompetence, denial, insincerity, and so on. On the other hand, given that in desire, when attending to the objects of one's desires nothing is actually represented as *obtaining* or as being *the case* or as objectively unfolding before one (when the object is an event), the fact that perception might present something as *not* being the case will have no bearing on whether one ought or ought not to desire that it *be* the case. In other words, one will not be caught up in any contradiction in simultaneously perceiving that *not-p* and desiring that *p*, even though in so doing one is representing both *not*-p in perception and *p* in desire, since p is *not* in desire represented as actually *obtaining*, ie. as an objective state of affairs, but only as a possible way the world *could* be, or perhaps additionally as a way the world is *to be made to be*. There being no contradiction here however would remain somewhat mysterious if one assumed that desiring that p involved (like believing or perceiving that p) having p in mind as an objective state of affairs. If this were so, in desiring that p while at the same time representing that *not-p*, one would in effect be representing p as both being the case and as not being the case simultaneously, thus landing one in a

[1] See for instance (Smith 1987) and (Sobel and Copp 2001).

contradiction of the very kind involved in simultaneously representing that *p* in belief and *not-p* in perception.

From all that has been said, it now clearly seems that episodes of representing the world (or it's directly presenting itself) as objectively one way or another cannot (at least in the case of the desires considered) capture the first-order perspective of desire, and might in fact be said to downright *contradict* some of the implications of the way in which the world features in our minds from our point of view in desire. Episodes of representing the world as objective can thus not be thought of as the first-order episodes that could potentially constitute the direct non-inferential grounds for our authoritative self-ascriptions of desires.

Having said that, a point in common remains between beliefs and certain desires, or between certain conscious manifestations of each. Both in consciously manifested belief *and* in consciously manifested desire, something (ie. that which one believes and that which one desires respectively) is being considered in one's mind or represented. The difference is in *how* these things are represented. That is, one *does* in consciously manifested desire (ie. when attending to the objects of one's desires) represent or have somehow in mind objects other than oneself (be they things, events, or states of affairs). Consciously attending to something other than ourselves or 'having something in mind' or 'being consciously occupied with something' is certainly part of what is involved in having a consciously manifested desire, and not just part of what is involved in having a consciously manifested belief. Having a desire is not, that is, just being in some kind of pure 'sensation' state. Specific episodes of representing things (possible events, possible ways the world could be, etc.) in certain ways may therefore still be found to constitute conscious expressions of desires, and so to constitute the potential immediate non-inferential grounds for self-ascribing desires. The problem of course is that in order for such episodes to constitute immediate grounds for mental self-ascriptions of desires, the way the objects of these desires are represented in these episodes must first of all *not* be as objective (for the reasons discussed above), yet still somehow such as to presuppose self-consciousness. So far though, the only sense we have been able to make of the idea of a world-directed state being at the same time a state of implicit self-consciousness, is by appeal to the world's striking us in having this state as *objective*. We are thus faced with a dilemma:

In order to move forward, we are going to have to either uncover some aspect of the way in which the world is apprehended in desire which does not involve objectivity yet which presupposes self-consciousness, or, we are going to have to settle for the view that, unlike cognitive states and perceptual experiences, desires are *not* actually states which we are able to self-ascribe directly on the basis of having them. Both of these options will be explored. Before turning to either though, there are a wide range of motivational states which we have so far overlooked, and which actually seem to go against the conclusions reached here that representing the world as being objectively one way or another cannot be part of how motivational states manifest themselves at the first-order level in consciousness. So far in fact, we have only considered a particular sub-class of desires and not yet any other emotions.

(2) Many emotions, and indeed a number of desires, *do* seem to manifest themselves (in part at least) in our coming to represent certain things as being objectively the case, and indeed sometimes in our coming to represent the very objects of our emotions and desires themselves as objective, ie. as objectively occurring events, as objectively obtaining states of affairs, or as actual mind-independent objects or people out in the world. That is, there do seem to be a number of emotions and desires which are directed towards objects, events, or states of affairs, actually represented as *objective*. In fact in some of these cases, representing the world as being objectively one way or another seems to not just be part of, but an *essential* part of how the world figures in our minds from the perspective of the particular emotion or desire in question.

It might be said for instance that it is essential to and *constitutive* of being paranoid that one come to represent someone as actually being out to get one. Or, one might argue that it is *essential* and *constitutive* of feeling guilty, say, that one represent oneself as actually having done something wrong, or that one represent that which one feels guilty about as having actually occurred. Similarly, in the case of many emotions directed at specific objects, events, or states of affairs, it might be said to be essential to one's feeling these emotions that the things in question be represented as mind-independent objects, as actually occurring events, or as objectively obtaining states of affairs. Consider for instance a case of loving a particular person, or fearing a particular spider, or being afraid of an unfolding event, say, lava flowing towards one, or being unhappy about some state of affairs. It would seem to be an essential part of having these emotions, that one be, when attending to

their objects, representing them as *objective*, that is, the person as actually existing, the spider as being a mind-independent creature before one, the lava as truly flowing, and the state of affairs one is unhappy about as objectively obtaining. If one did not have the spider in mind as actually before one objectively but merely as something that *could* be present, or if one did not represent the state of affairs one is unhappy about as actually obtaining, it is doubtful that one would be quite so afraid, or at all unhappy. Thus, in many cases it seems that representing the objects of one's emotions and desires as objective, or representing various things as actually being objectively the case (as in the examples of guilt and paranoia) *can* be an essential and *constitutive* part of the way in which these emotions and desires manifest themselves at the world-directed conscious level.

But now, if in being afraid of something, say, of an oncoming object, or if in having a pang of desire, say a sexual desire for someone, one's attention is focused out on the world, a world which clearly appears to one a certain way, or is represented as being a particular way, in fact amongst other things as *objective*, there should it seems be no real problem for fitting at least *these* emotions and desires into the account of self-knowledge already put forward for cognitive states. Could one not, that is, just appeal to the self-world dualism implicit in representing the world as objective to account for how these conscious manifestations of emotions could be states of implicit self-awareness, and so direct grounds for mental self-ascriptions of the relevant underlying emotions?

In a sense it is true that such episodes of representing the objects of one's emotions as objective could be shown to be states of implicit self-awareness. The more fundamental problem however would still remain of accounting for how such states could be states not just of implicit self-awareness, but of implicit awareness of oneself as having a particular *emotion* or *desire* and not just a cognitive perspective on their objects. To see how this problem arises, one need only look back at how objectivity did the trick for cognitive states and perceptual experiences.

Representing something as objective, in the relevant sense of 'objective', is to be representing it as mind-independent, that is, in effect as independent from a point of view one has upon it. In representing the world as objective therefore, one is also thereby implicitly representing oneself as *representing* it or *perceiving* it, that is, implicitly representing oneself as having a cognitive or perceptual point of view upon it. Objectivity however only implies this distinction between the world and one's

cognitive or perceptual point of view upon it; it does not imply an *emotional* relation between oneself and the world. Representing therefore an object of fear or a person one desires merely as objectively before one, or a state of affairs one is unhappy about as objectively obtaining, cannot, it seems, even if such representations are constitutively tied to having the emotions in question, be a source of knowledge of oneself as being *afraid* or of oneself as *desiring* or of oneself as being *unhappy* and so on, but only of oneself as *representing* or *perceiving*, that is, of oneself as having a detached, purely cognitive or perceptual point of view upon these things. In fact, if particular emotions were to manifest themselves *only* in our coming to represent various things as being objectively the case – eg. someone as being actually out to get us, or a person as having various negative character traits, etc. – it is unlikely that we would find ourselves inclined, merely upon being asked the question, to self-ascribe an emotion (eg. paranoia, hatred, etc.) rather than to make a primarily evaluative judgment about the person in question, their intentions, or the situation as a whole.

The main problem thus seems now to be *not* that having consciously manifested emotions might not involve at the personal level representing the objects of our emotions as *objective*. Instead, the problem is that the objects of our emotions striking us as objective or being represented as objective even if crucially involved in having certain emotions, is still not enough to ground immediate knowledge of ourselves as having the emotions in question as opposed to merely having a non-emotional point of view on their objects. If specifically *emotional* self-consciousness is to be shown to be already implicitly present at the first-order level, emotions will have to be shown to contribute something additional to our conscious first-order perspective than our representing their objects as objectively existing, occurring, or being the case.

One first appealing attempt at this might be extracted from cognitivist theories of emotions. This attempt will ultimately be shown not to work. Seeing *why* however will prove invaluable in allowing us to narrow down the requirements on a more successful account.

6.3 The cognitivist proposal

Put briefly, according to cognitivist accounts of emotions, to have an emotion is essentially to be making a special kind of judgement, or more generally, to be

representing (with or without active assent) the object of one's emotion as having some special emotion specific property such as that of being desirable, attractive, frightening, dangerous, worthy of suspicion, worth possessing, worth avoiding, and so on.

Applying this idea to the problem of self-knowledge, it might be said that we are able to know what emotions we have directly on the basis of having them, in virtue of the emotion-specific content of these instances of representation that are constitutive of our having these emotions. Or, if one does not want to endorse all aspects of the cognitivist theory (ie. that emotions are nothing more than such instances of representation), one could treat these episodes of representing the world as having emotion specific properties merely as the *expressions* of emotions in consciousness, and suggest that we are at least able to *self-ascribe* the emotions specified by the contents of these conscious episodes (even though there may be more to emotions than these expressions) directly on the basis of having these first-order episodes in which our emotions are manifest. Now, at first glance at least, this approach to the problem of authoritative self-knowledge for emotions seems to have much going for it.

To begin with, it provides, as needed, an account of the contribution that emotions might make to how things are for us at the *first-order* (ie. *world*-directed) personal level. Next, it also seems to tell us something about how things are at this first-order emotional level which already makes reference to the having of emotions (eg. reference to desire in things being represented as 'desirable', reference to fear in things being represented as 'frightening', etc). A further advantage of adopting this approach to the problem of self-knowledge for emotions is that one can do so it seems without having to take on many of the more problematic claims made by defenders of cognitivist theories, that is, claims about what emotions actually *are* or consist in and claims which are generally brought in to solve problems having nothing to do with self-knowledge.

Proponents of the view that emotions are a special kind of judgement or cognition,[1] are in fact generally not motivated by considerations to do with self-knowledge, but by considerations primarily to do with questions about the *rationality* of emotions and their role in the rationalisation of other states and actions. Taking

[1] See for instance (Solomon 2003) and De Sousa (1987).

emotions to be kinds of judgements may in fact make it easier for us to see how they could be thought of as rational or appropriate in particular circumstances. If fear for example involves representing the object of one's fear as being dangerous or threatening, being afraid could be said to be 'appropriate' in certain situations if indeed danger looms, and 'rationally grounded' if one perceives something to be dangerous. Similarly, taking emotions to be kinds of judgements would make it easier to see how having a certain emotion might make following a particular course of action more appropriate than following others. For instance, if being afraid involves representing the object of one's fear as 'worth avoiding', avoiding the object would come out as being the rational thing to do when afraid.

Having said this, the view has been widely criticized, and in many cases rightly so. Without going too much into the details of these criticisms, there does seem to be more to having most emotions than making evaluative judgements. Two points are particularly damaging to the cognitivist proposal. First, one could it seems perfectly well be judging something to be, say, dangerous or frightening or 'worth avoiding' yet not be afraid of it, or, judging someone to be, in some objective sense, desirable or attractive without desiring them. Conversely, one could it seems also be afraid of something while *not* judging it to be frightening, or be attracted to someone despite *not* taking them to be objectively speaking attractive. Clearly therefore there must be more to having these and other emotions than making such judgements. Furthermore, there seems to be nothing about judging something to be generally speaking desirable or attractive or worth having, etc. which could explain why when desiring something, but not when making the above types of judgements, one tends to feel directly drawn into action, and drawn specifically either towards the object of one's desire or into changing the world to match one's desire's content. In other words, the directly motivational aspect of emotions and desires cannot be accounted for via a purely cognitivist theory of what emotions are either.

Fortunately, one might say, the problems that arise for the cognitivist theory of emotions tend to apply to it mainly when considered as a theory of what emotions *are* or of what having an emotion *consists in.* Adopting the cognitivist idea purely for the purposes of solving the problem of self-knowledge need not involve making any such strong claims. All that is needed from the cognitivist proposal to account for the possibility of authoritative knowledge of our own emotions, is for it to identify some (possibly contingent) feature of a certain range of our emotions, or of their

manifestations in consciousness, which would immediately reveal to us from our first-order perspective the fact that we have them. This aspect need not be exhaustive of their conscious manifestation; it certainly need not be what having an emotion *consists* in, nor indeed what an emotion's being *conscious* consists in.

A useful analogy can be found in thinking about the case of belief. Beliefs can be thought of as manifesting themselves in our conscious minds as judgements, without it following from this that making such judgements or even being constitutively inclined to do so is exhaustive of what having a belief involves or consists in. Nonetheless, it is essentially through the manifestation of our beliefs in conscious acts of judgement that we are able to *introspect* them. Something similar might be true of emotions. What we are looking for is some aspect of the conscious manifestations of emotions at the first-order (ie. world-directed) level which could itself directly reveal to us that fact that we have them, but which, as mentioned, need not be exhaustive of what having an emotion involves whether at the conscious level or more generally speaking, let alone what having it actually *consists in*. To this extent then, appealing to the cognitivist idea merely to solve the problem of self-knowledge need not be subject to the various criticisms that might go with the idea that emotions just *are* kinds of judgements or cognitions, or episodes of representing things as having properties of desirability, frightfulness, and so on. Nonetheless, upon closer examination, parallel problems to those that arise for cognitivism *per se* carry through to this more restricted appeal to it.

The main problem with appealing to the cognitivist idea in order to solve the problem of self-knowledge for emotions lies in the fact that it is not at all obvious that something's presenting itself to one, or being represented by one, as having general properties such as of desirability, frightfulness, worthiness of respect, blameworthiness, etc. actually *does* presuppose reference to oneself as truly desiring, fearing, feeling respect for, etc. the things/people so represented. And, if it does not, it cannot provide us with direct evidence that we have the relevant desire or emotion. As mentioned above, something's being desirable or frightening or worth avoiding does not in any way guarantee or entail that one feels any emotion towards it. Things may just happen to strike one as having these general properties. Someone can appear to one as having many qualities, including as being lovable and attractive, without one actually loving them or being attracted to them. Similarly, a situation can strike one as being generally speaking dangerous or frightening without one actually being afraid in

that situation. One's representing therefore something as having such general properties cannot only not be what having an emotion *consists* in, but it can also not it seems provide a sufficient ground for immediately self-ascribing an emotion. The apparent desirable/ frightening/ etc. character of an object from our point of view does not tell us that we *desire* it or *fear* it.

In sum, the cognitivist proposal seems ultimately to fail not only as a theory of what emotions *are* but also as a theory of the conscious manifestation of emotions in consciousness, at least insofar as these conscious manifestations are supposed to provide the immediate grounds for knowledge of our own emotions. Put differently, the cognitivist proposal seems to fail as a way of accounting for the possibility of implicit emotional self-consciousness when attending to the *objects* of one's emotions. Moreover, this proposal seems to fail for essentially the same reasons as did the attempt to account for the possibility of implicit emotional self-consciousness simply by appeal to the fact that emotions and desires involve (amongst other things) representing their objects as *objective*. What the cognitivist theory has brought out more clearly though, is that if emotions and desires are to be shown to manifest themselves in consciousness in such a way that involves being at the same time implicitly aware of oneself as having them, the objects of these emotions will have to be shown to strike us, from our first-order conscious point of view, as more than just objective things, events, or state of affairs, bearing various general properties, whether evaluative properties or otherwise.

In sum, the source of our worry now lies not in the thought that the account of pre-reflective self-consciousness put forward for the case of cognitive states is unable to account for our being implicitly *self-aware* in having an emotion or a desire, but in the thought that it is unable to account for our being implicitly aware of ourselves as having an *emotion* or a *desire*. If it is in virtue of how we represent the world as being or in virtue of how it appears to us to be that we are self-aware in desire and emotion, the worry is that *unless* the way things appear to us or are represented as being in desire and emotion is radically different from the way things appear or are represented in judgement and perception, all we can be aware of in having an emotion is of ourselves as *representing* and *perceiving* but not specifically of ourselves as *desiring* or *fearing*, etc. This problem is likely to persist until desires and other emotions are shown to contribute something to the way things appear to us at the personal level, which implies reference not just to the fact of our having a point of view on some

aspect of the world but also to the fact that we have an *emotional* point of view upon it. The attempt just made in this direction by appeal to the cognitivist theory of emotions has failed. All that representing something as *attractive* or *desirable* or *dangerous* or *worth avoiding* seems able to ground is awareness of ourselves as representing or perceiving it as having certain general properties, but not of ourselves as *desiring* or *fearing* it, regardless of whether we take it to be objectively speaking attractive or dangerous. Nothing about representing things as having the kinds of properties posited by the cognitivist in other words seems to imply reference to oneself as having emotions. Cognitivist accounts of what is involved in emotion and desire at the personal level can thus not be the whole story, and for more reasons than one.

Cognitivist accounts cannot be the whole story of our first-person conscious perspective in having an emotion or desire not just because they are unable to make room for implicit awareness of our emotions and desires, but also because they are equally unable to accommodate certain other differences that clearly exist at the personal level between having an emotion and making a judgement. As mentioned earlier, there seems to be nothing about representing something as being generally speaking desirable or attractive or worth having that could explain why when we have a desire or emotion we are directly drawn into action, and in particular that could explain why we directly *feel* compelled or drawn towards very specific things. A correct theory of the manifestation of emotions and desires in consciousness must in the end be able to provide an account of the contribution that emotions and desires make to how things appear to us from our own point of view such that (a) having these emotions and desires is already presupposed by the way things appear to us at the first-order (ie. world directed) level, and (b) such that our feeling directly moved to act (and not our just 'finding ourselves' acting) is made intelligible. For present purposes, it is of course the first question that is most pressing. The second one however helps emphasise (and will end up, we will see, being answered by a solution to the first) that the way things appear to us in having an emotion or a desire must be such as to directly imply more than just having a point of view on an objective world of items bearing various general properties.

With these considerations in mind, it is time to move on to put forward an account of the manifestation of emotions and desires in consciousness which will be able to account for all these distinctive features of emotions, rectify mistaken

assumptions about how things appear to us at the first-order level in having such states, and thereby ultimately rid us of the nagging sense that our awareness of our own emotions and desires cannot be accounted for along a similar model to that adopted for the case of judgements and perpetual experiences, that is, along a model based on the idea that having states of the kind we are able to introspect and self-ascribe with authority, involves being already at the first-order level somehow implicitly aware of ourselves as having them.

6.4 The way forward

Without abandoning the basic idea behind our present solution to the puzzle about mental self-ascription (ie. the idea that it is in virtue of our self-ascribed states' being somehow states of implicit self-awareness that we are able to immediately self-ascribe them directly on the basis of having them), we might it seems be able to find a solution to our puzzle for the case of mental states other than beliefs and experiences as of an objective world by appealing to some of the *other* aspects of the way in which we think about and experience the world at the first-order level, aspects which might, like the phenomenological objectivity of our experiences for the case of cognitive states, presuppose our being implicitly aware of ourselves as having distinctively *emotional* attitudes towards the world.

Doing this however will require two things. First, it will require as already mentioned putting forward a more precise account of how emotions and desires manifest themselves at the conscious world-directed level. That is, an account of the world-directed phenomenology of emotions and desires will be needed, ie. of the distinctive way in which the objects of our emotions present themselves to us, from our point of view, when we consciously attend to them. Next, a very specific aspect of this world-directed phenomenology of emotions and desires will need to be uncovered (ie. at least of those emotions and desires we are able to know directly on the basis of attending to their objects), an aspect which presupposes *self-consciousness*. Doing the latter will come with the greatest challenge.

First, as we have already seen, we will *not* be able to appeal to the phenomenological objectivity of our experiences, episodes of visualisation, etc. alone. Instead, a feature of the world-directed phenomenology of emotions and desires will need to be uncovered implicit in which is already the fact that we have an *emotional*

attitude towards the world and not just a cognitive or perceptual point of view upon it. Second, the account we must arrive at of how we are able to know our own emotions will have to be such as to still make room for the difficulty of access and openness to interpretative error characteristic of our knowledge of many of our own emotions. Even when our knowledge of our own emotions is based on *world*-directed states rather than self-directed ones, reaching it tends to be a more difficult process, and one more highly susceptible to error than that of attaining knowledge of our own cognitive states and perceptual experiences. This will have to be made room for. Recall for instance the example mentioned earlier, of your coming to realise that you have strong feelings towards someone only after observing them for a while, thinking about them, listening to them, recalling past events involving them, etc.[1] Although all of your reason-giving states are *world*-directed in this example, the knowledge of your own emotion for which they provide a ground is neither instantaneous nor largely immune to error. In brief, the account of self-knowledge to be put forward here will have to be able to accommodate the features that our knowledge of our own emotions shares with the knowledge we have of our own cognitive states *without* doing so in a way that renders the differences between the two kinds of knowledge thereby unintelligible.

The next two chapters will look at both these similarities and differences. Starting with the similarities, the first question that will need to be addressed is that of how attending to the *world* can give us knowledge of our own *emotions*, as it can of our own cognitive states. Towards this aim, an account will be needed of how our emotions and desires manifest themselves at the world-directed conscious level, focusing specifically on how they might do so in a way that already presupposes awareness ourselves as having an emotion. That is, if our puzzle about direct reason-based mental self-ascription is to be solved for cases of self-ascriptions of emotions, our earlier investigation of how the world strikes us in experience, thought, imagination, etc., will first have to be delved into further, with particular focus this time on the phenomenology of world-directed consciousness when in the grip of a desire or emotion, or, less episodically, when having an underlying emotional disposition towards some aspect of the world (eg. a general fear of spiders, love for one's family, etc.). This is where the next chapter will begin.

---

[1] See p.130 above

Before embarking on this task however, it should be borne in mind that what we will need is *not* a full account of what emotions and desires are, nor of what an emotion's being conscious consists in, nor indeed a full account of every aspect of the world-directed phenomenology of emotion. Rather, what we will need to find is a (possibly contingent) feature of the way in which certain emotions (ie. those we are able to introspect) manifest themselves in first-order consciousness, such that self-ascribing them directly on the basis of attending to the world is possible. The time has come to look at this, ie. to delve further into how emotions and desires might indeed manifest themselves or find expression at the first-order level.

Chapter 7: Expressions of emotion and desire

Chapter 6 has brought to the foreground a problematic feature of our knowledge of our own emotions, namely that in a wide range of cases, in self-ascribing our own emotions we turn our attention primarily to the *world* or to the *objects* of our emotions rather than to any explicit data regarding ourselves, such as to our behaviour, to our current sensations, or to any other of our mental states. This, we have seen, seems moreover to be not just something we *do*, but generally something that *makes sense* to us from within our own outward-looking perspective. We do not, that is, just 'find' ourselves self-ascribing emotions on certain occasions when turning our attention to the world, but rather, as with cognitive states, our doing so on these occasions seems somehow *appropriate* to us. We seem to turn our attention to the world for *evidence* or *confirmation* that we do or do not have a particular emotion.

How can this be? The fact that we often proceed in this manner has raised what is now a familiar explanatory challenge: how can attending to the *world* possibly give us knowledge of a certain class of our *mental states*, in this case of our *emotions*? This challenge has come this time however with an added complication. We have seen that the objectivity of the world (as experienced) from our point of view cannot alone solve the problem for the case of emotions as it did for cognitive states. Some other aspect therefore needs to be uncovered, specific to the way emotions manifest themselves at the world-directed level in consciousness, implicit in which might already be the fact that we have an *emotional* attitude towards the world and not just a cognitive or perceptual point of view upon it. The aim of this chapter will thus be to do just that. An account will be put forward of the world-directed phenomenology of

emotion, that is (as will be explained) of the way our emotions manifest themselves or are 'expressed' in our world-directed conscious experiences, such that authoritative knowledge of these emotions, based directly on looking out at the world, is possible.

With this aim in mind, the chapter will begin by first making clearer sense of the ideas of emotions 'manifesting' themselves or being 'expressed' in consciousness, and in particular of their doing so in our *world*-directed conscious experiences. Having done that, this chapter will argue, first of all, that our emotions *do* manifest themselves at the world-directed level, and that they do so essentially by emotionally 'colouring' the world (or certain aspects of it) from our perspective, that is, as will be explained, by making us come to experience the world in a manner that is 'expressive' or 'evocative' of these emotions. Then, by appealing to considerations about the distinctive phenomenology of our emotionally coloured experiences, as well as to recent discussions and data from the psychological literature, it will suggest that and how our emotions (or at least those that we are able to know with authority) do not just *tinge* the world from our perspective, but do so in colours (a) that are implicitly self/world *relational*, and (b) implicit in which is already a distinctively *emotional* or *motivational* self/world relation; not just a cognitive or perceptual one.

First though, we need to get clear about what exactly it is for emotions to 'manifest' themselves or to be 'expressed' in our world-directed conscious experiences. And, to this end, it will be worth first establishing what it might be for emotions to find expression in more familiar circumstances, that is, not in consciousness, but in outward behaviour. It will be suggested here that in a somewhat analogous way to the way in which we may be able to read off other people's mental states directly from the *outward* manifestations of these states in their behaviour, we might also be able to read our *own* mental states (and emotions in particular) directly from their '*inner*' manifestations, not in behaviour, but in first-order consciousness, that is, in our ways of experiencing the *world.*

7.1 Outward expressions of emotion

Consider the following concrete examples. Outward expressions of emotion can it seems be anything from facial expressions (eg. smiling, frowning, having a certain look in one's eyes), to other spontaneous behaviours such as crying, laughing, hanging one's head, dragging one's feet, jumping up and down, punching the air,

letting out sighs of relief, exclamations of surprise, and so on, all the way to (though more controversially) quite elaborate intentional actions such as putting pins in a doll resembling someone one dislikes, or tearing the eyes out of a photograph representing someone one is angry at.[1] Additionally, emotions can it seems also be expressed in behaviours and actions that are not *themselves* 'expressions' of emotion, but that are nonetheless executed in some *manner* that is 'expressive' of an emotion, as when opening a door hesitantly for instance, or playing a piece of music with feeling. In other words, emotions can on the face of it be expressed in a variety of ways and in a wide range of behaviours and acts, ranging from automatic nervous system reactions to quite elaborate symbolic intentional actions, and even to particular *ways* in which some of these behaviours and actions are carried out.

The differences between these examples however are significant. In amongst other ways they differ (a) in whether the supposed outward 'expression' is an intentional action or not; (b) in whether the emotion is supposed to be expressed in the behaviour *itself* or only in the *manner* in which it is carried out; and most of all (c) in whether in each case the situation should be thought of as one where a *particular* emotion (ie. an emotion one *actually has*) is being 'expressed' in a specific instance of behaviour or instead as a case where the behaviour, independently of whether one has an emotion or not, is 'expressive' of a *type* of emotion. Some of these examples could it seems even be taken both ways, but not all. If any theoretically utilizable notions of 'expression' or 'expressiveness' are thus to be extracted from these cases and used in solving our problem about self-knowledge, a number of distinctions will first need to be drawn, and a number of questions addressed separately. In what follows, three questions in particular will be taken in turn:

(1) What is it for a *particular* emotion (ie. an *instantiated* emotion; a state one is actually *in*) to be 'expressed' in an instance of behaviour? Or, to put things the other way around, what is it for an instance of behaviour to constitute a direct 'expression' of a particular emotion one now has? (2) What is it for an instance of behaviour to be this time 'expressive' of a *type* of emotion? That is, what is it, say, for a certain 'look in one's eyes', or for a certain way of playing a piece of music, to be expressive of anger or of sadness or of joy? Is it just for the behaviour to be, as above, an *expression* or direct *manifestation* of an *actual* emotion one now has? Or, could a

[1] This example is taken from Goldie (Goldie 2000, ch5) and will be discussed in further detail below.

certain instance of behaviour (let alone a *type* of behaviour) count as emotionally *expressive* or *evocative* of a type of emotion without actually being the outward expression of anyone's present mental state? Finally (3), the more specific question will be worth addressing (even if only to further clarify our answers to the first two) of whether it is ever appropriate (and if so, in what sense) to think of intentional actions (such as some of those cited in the examples above) as being either instances of 'behavioural expressions' or forms of 'expressive behaviour'. Concerning this last issue in fact, it will be suggested that the controversy surrounding it arises in part from a failure to disambiguate between 'behavioural expressions' and 'expressive behaviour'. Let us therefore begin with the first two questions, taking them in turn.

What are 'expressions' of particular *emotions?*

Paradigm cases of 'expressions' of one's emotions or of direct 'manifestations' of emotions one now has include, most uncontroversially, behaviours such as smiling spontaneously when happy, bursting out laughing when amused, automatically raising one's voice as one's anger increases, and so on. In other words, standard cases of expressions of emotion tend to be episodes of essentially spontaneous, non-intentional behaviour, which, one might add, is not surprising since it would seem to be of the very essence of something's being an *expression* of a mental state as opposed to an action *out of it*, that it not be something done on purpose, ie. that it not be an intentional *action.*[1] As soon as smiling, to take an example, becomes intentional, it seems to lose its whole status as an *expression* or *manifestation.* Consider for instance the difference between your smiling spontaneously upon being told some good news, and being happy *and* smiling, say, in order to let others know that you are happy. In both cases your facial expression

[1] Not all actions, it has been argued, are intentional (see in particular O'Shaughnessy's discussion of sub-intentional action in O'Shaughnessy 1980). An example of a sub-intentional action might be for instance a case of flicking around a pen in one's hand while, say, thinking about something else. Could such an action ever be an 'expression' of emotion, say, of nervousness? I would tentatively say that it can, as can other complex yet spontaneous behaviours such as punching the air or jumping for joy. The important contrast when distinguishing possible 'expressions' from 'non-expressions' is I believe (in part at least) one between *intentional* and *non-intentional* behaviour from emotion. For present purposes therefore, cases of sub-intentional action can be put on a par with purely reactive behaviour. Of course, there may be borderline cases even between intentional and sub-intentional actions, thus making it remain unclear in some cases whether a particular instance of behaviour should be thought of as an 'expression' of emotion or rather as an action *out of* emotion, ie. as an action *motivated* by an emotion. The mere existence of borderline cases does not however in itself invalidate the meaningfulness of a distinction, and so will not be dwelled upon here.

signals to others that you are happy, but in the latter case your happiness is not revealing *itself* in your behaviour, but rather *you* are revealing *it* by behaving in that way. That is, you are smiling essentially *in order* to signal that you are happy, rather than your happiness itself directly 'coming through' or 'finding expression' in your behaviour. Behavioural 'expressions' or 'manifestations' are in other words things one to some extent just 'finds' oneself doing, often even in situations where one might have preferred not to.[1]

None of this is to say of course that intentional control over one's emotional outbursts is not to some extent possible. Clearly it *is* possible. One can purposely suppress one's expressions of emotion in certain situations and purposely *not* suppress them in others. The only *actions* carried out though in such situations are those of *suppressing* or *not suppressing* one's behavioural impulses, not those of intentionally *manifesting* or not one's mental states. Once one stops suppressing an expression of emotion and allows the emotion to 'come out' so to speak, the actual *coming out* of the emotion in one's behaviour is not itself an intentional action, just as, conversely, the act of allowing one's emotion to come out is clearly not an expression of emotion. Consider for example a situation where you finally burst out laughing uncontrollably after having resisted this impulse for several minutes while stuck in a situation where such behaviour would have been unacceptable. Your action here is that of finally *allowing* your emotion to manifest itself, rather than that of you *manifesting* it, if indeed saying the latter makes any sense at all. 'Expressing' or 'manifesting' an emotion is it seems clearly not something that you purposely *do*.[2]

So, what follows from all this? Having noted the essentially non-intentional character of genuine cases of 'expressions' of emotion, one thing that immediately seems to follow is that some of the examples of 'expressions' mentioned at the very beginning, can no longer strictly speaking be counted as such. This includes in

---

[1] The term 'expression' is actually ambiguous in that there is a sense in which one can intentionally 'express one's fears' or 'express one's doubts' where this means essentially *putting into words* one's fears or doubts, in the same way that one might put into words someone *else's* feelings. This is clearly not the sense of expression at issue here. In fact, in the second sense, one could it seems 'express one's fears' yet do so in a completely controlled and unemotional manner, that is, while letting no emotion transpire or be expressed in one's behaviour in the present sense.

[2] I have been using the terms 'manifestation' and 'expression' interchangeably here, though I do believe that there is a significant distinction to be drawn between them. This distinction will be spelt out in more detail when considering the notion of 'expressive behaviour' (in contrast to that of 'behavioural expression'). Very briefly though, the idea will be that 'expressions' are essentially a *sub*category of 'manifestations'. They are what one might call 'expressive' or 'emotionally coloured' manifestations of emotion. Whether emotionally 'coloured' or not though, outward manifestations remain non-intentional states. The distinction therefore between manifestations and the narrower category of 'expressions' is not of immediate relevance here. It will be so further below.

particular the examples of symbolic intentional actions out of emotion such as the one borrowed from Goldie, ie. that of tearing the eyes out of a photograph representing someone one is angry at. Insofar as this is a *motivated* action, one carried out *because* (ie. for the *reason* that) one is angry, it can no longer be thought of as a genuine case of 'expression' on the present account. What should we make of this result? And, most importantly, if it is (as it seems) so obvious that the phenomenon of 'expression' is by its very nature a non-intentional one, why might it intuitively seem to us nonetheless, as it does to Goldie, that such actions *should* in some sense still be counted as 'expressions' of emotion? As said earlier, the specific case of such actions will be examined in further detail once all the distinctions between 'expressions' and 'expressive behaviour' have been drawn. It is however worth noting immediately that, as far as what has been said so far goes, intentional actions out of emotion have only been ruled out from being 'expressions' of emotion, but not yet from being potentially 'expressive' forms of behaviour. It may thus still turn out that they can be the latter. Not being an expression of an emotion, we will in fact see, does not preclude an action from being an *expressive* action, which may well salvage much of the substance of what Goldie wants to say about these cases, despite his referring to them interchangeably as 'expressions' of emotion and as 'expressive' actions.[1] More on this below.

So far, we have seen (i) that outward expressions of emotion are essentially *non*-intentional episodes of behaviour; and (ii) that intentional actions therefore, even those motivated by an emotion cannot count as genuine instances of 'expressions' of emotion though they may still turn out to be emotionally 'expressive' actions in some sense yet to be specified. What about some of the other examples listed earlier of emotions being supposedly expressed, not in performing an intentional action itself, but nonetheless in the *manner* in which an intentional action is carried out? These, it seems, despite being in a sense manifestations of emotions in *actions*, will *not* be ruled out from being genuine cases of 'expressions' on the present account. From the fact that the phenomenon of 'manifestation' or 'expression' is not a reason-guided phenomenon, it does not in fact follow that all instances of behaviour *in the course of which* emotions are expressed or manifest themselves cannot be reason-guided. The behaviour itself (ie. the fact of behaving at all, as well as the performing of a

[1] See (Goldie 2000 ch. 5)

particular action rather than another) may be intentional (eg. talking to someone), yet one's performing it in a certain *manner* (eg. very loudly) may be to some extent *non-*intentional, and to that extent potentially a direct outward 'manifestation' of an emotion. To go back to one of our initial examples, that of opening a door hesitantly, a person S's action of opening a door might be perfectly intentional and reason-guided – S might be opening the door because she needs to get something on the other side – yet this action might at the same time be influenced, non-rationally and non-motivationally, by an emotion of hers, eg. fear. That is, S might be afraid of what awaits her on the other side, and her fear might thus spontaneously manifest itself in her way of opening the door (ie. slowly and hesitantly rather than swiftly and with confidence), a way in which she does not specifically *intend* to be opening the door, yet an aspect of her action which is clearly determined by her fear, and so a possible direct 'expression' or 'manifestation' of her fear. What is the importance of this?

In the second half of this chapter, it will be shown that in a parallel manner, when emotions manifest themselves in *consciousness* (as opposed to in behaviour), they do not do so only by *giving rise* to particular conscious episodes (in the way that they give rise to particular behavioural episodes), but also on occasion by altering or transforming conscious states one already *has*, or *would have*, independently of having any emotion. For example, one might be recalling a past event because one has been asked to recount it (ie. not because one feels any particular emotion), yet the *way* in which this event might be remembered, or experienced by one in memory, might be to an extent transformed by how one feels about it (eg. embarrassed, proud, etc.). Put more generally, emotions will be shown to manifest themselves in consciousness not just by causing specific conscious episodes, such as by making us attend perceptually to certain objects rather than others, by making us recall particular past events, or by making us imagine certain types of situations,[1] but also often by affecting the *way* in which these objects, events, and situations are experienced, remembered and imagined from our conscious point of view.

Sticking for the moment though to the more familiar cases of expressions of emotions in *behaviour*, the point to note for present purposes is that, outwardly at

[1] That emotions direct our attention, and determine the salience of certain things rather than others in perceptual experience, memory, etc. is a point often made in the psychological and neuropsychological literature on emotions. See in particular (Damasio 1994) and (De Sousa 1987). My fear of rats for instance might result in my attention being drawn to such creatures, or to the possibility of encountering such creatures, when, say, walking through an underground passage.

least, emotions find expression not only in involuntary episodes of behaviour, but also in the involuntary *ways* in which such behaviours (and often independently motivated voluntary actions) are carried out. Either way though, the phenomenon of expression remains a strictly non-intentional one. But, granting now that this is so, ie. that the relation of 'expression' is never a reason-giving one even when it holds between emotions and ways of executing reason-based actions, the question arises of what kind of a relation it *is* then.

A first thing to remember is that expressions of emotion are always expressions of *particular* emotions, ie. of states one actually *has* – an emotion can hardly manifest itself in one's behaviour unless one actually has the emotion in question. The relation between emotions and outward expressions is therefore going to have to be one that can hold between two concrete existences. Yet, since we have seen that it will not be a *reason-giving* relation, nor for that matter a purely coincidental one (recall, simply being happy *and* smiling is not enough for one's smiling to constitute an expression of one's happiness), the obvious alternative is that it is a purely *causal* relation. Thus, we might say: for an instance of behaviour to constitute an 'expression' of emotion, it must be directly *triggered* (in a purely physical and non-motivational sense) by an emotion one now has.[1] Is this enough though? Can being just directly physically caused by an emotion be all there is to being an 'expression' of emotion? Something about this proposal seems deeply unsatisfying.

To see this, consider the difference between the following two types of cases: (a) having a certain 'angry' look in one's eyes, or smiling, laughing, etc; and (b) twitching in some unusual way, or clumsily tripping up while in the grip of joyful enthusiasm or anger. In reality, instances of both of these types of behaviour could it seems be on various occasions directly caused by one's emotions. Yet only the former are behaviours of a kind we would generally be prepared to call 'expressions' of

[1] William James would argue that the causal relation should actually be thought of as going the other way around (See James 1884). On James's view, it is essentially visceral disturbances and other physiological changes (or more specifically one's *perception* of these) that cause emotions rather than vice versa. Emotions are then in turn to be understood essentially as sensations arising from such causes. Oddly however, within this category of visceral or physiological changes, James includes not just basic nervous system reactions, but also behaviours that we would more naturally classify as 'expressions' of emotion (ie. facial expressions in particular) thereby turning expressions of emotion into the *antecedents* of emotions rather than, as seems more intuitive, the *results* of them. James's view will be returned to briefly when discussing how emotions manifest themselves in consciousness. This aspect of it however (ie. the view that expressions of emotion are essentially the *causes* of emotions) will be left aside. It will simply be assumed that the causal relation (if any) between emotions and their outward expressions, goes from emotions to outward behaviour.

emotion – the latter being at best thought of as mere one-off consequences or incidental manifestations of them. The holding of a direct causal relation between a particular emotion and a particular bodily movement is thus clearly not sufficient for the latter to count as an 'expression' of the former. Moreover, upon reflection, it is not even obvious that the holding of a direct causally *constitutive* relation between one's having an emotion of a certain kind and one's behaving in a certain way would be enough for one's behaviour to constitute an 'expression' of this emotion. Physiological changes such as trembling for instance, although arguably systematically correlated at the physical level with having certain specific types of emotion (eg. fear), are not events we would generally think of as 'expressions' of fear. For an episode of behaviour to constitute an 'expression' of emotion therefore, there must be more to it than it's being simply *caused* by an emotion, whether this is a one-off causal connection or an instantiation of a more systematic causal correlation between emotions of that type and the relevant forms of behaviour. So, what is missing?

One way of bringing out what might be missing is by looking at cases of emotional pretence. When pretending to be happy or angry or sad one will generally tend to simulate certain types of behaviour rather than others. In particular, when pretending to be angry for example, one might try to give someone an 'angry' type of look or raise one's voice (ie. carry out behaviours of kinds which, if done spontaneously, would be thought of as 'expressions' of anger), while one would *not* in similar circumstances (ie. in pretending to be angry) generally do such things as simulate some odd twitch, or pretend to trip over a carpet (ie. carry out behaviours *not* generally thought of as 'expressions' of emotion, however much they might be caused on particular occasions by one's emotions). Now what this immediately suggests is that, at the very least, there must be something about behaving in the ways we think of as potential 'expressions' of emotion, to directly *fool* an onlooker into believing that one has the intended emotion. That is, there must be something about episodes of, say, smiling, laughing, shouting, etc. which makes us directly *read into* these behaviours happiness, anger, sadness, etc. in a way that we do *not* immediately do so into other behaviours, no matter how much they might *in fact* be caused (whether in isolated instances or more systematically) by these same emotions.

There are a number of things that might explain this, but the most obvious explanation (and I will suggest ultimately the correct one) is that in cases of

behavioural effects of emotions we *do* think of as 'expressions' in contrast to those we do *not* think of as 'expressions', there is an additional direct *conceptual* link between the *way* one is behaving (eg. smiling) and the type of emotion causing one to behave in that way (eg. happiness). Immediately revealing of this is in fact already our tendency to refer to many outward expressions of emotion as 'angry' looks in one's eyes, 'guilty' expressions on one's face, 'confident' ways of behaving, and so on. We seem to think of certain forms of behaviour in other words as being already themselves intrinsically of an emotional type, or as being themselves directly *indicative* or *evocative* of certain types of emotions. More than that, when someone smiles or looks at us in an 'angry' type of way, we often take this to directly *mean* they *are* happy or angry, in a way that we would not take anyone's twitching or trembling in itself to signify that the person in question actually has any specific emotion.

Positing the existence of a direct meaning relation between expressions such as smiling and emotions such as being happy, may not however be the only way of explaining the apparent immediacy of the mental transition we make between observing people behave in these ways, and conjuring up, or attributing to them specific emotions. One might protest that it does not follow simply from the fact that we are quicker to assume that someone is happy upon seeing them smile than we are to assume that they are angry upon seeing them twitch, that there is any more of a direct meaning relation between smiling and being happy than there is between twitching and being angry. It might just be that given past associations, we have come to have better reason to believe that someone is happy upon seeing them smile, than to believe that they are angry upon seeing them twitch, and so for this reason have acquired a tendency to very quickly infer that they are happy upon seeing them smile, but not so quickly to infer that they are angry upon seeing them twitch. In both cases though, in coming to the conclusion that the person behaving in the relevant way is happy or angry, we might be proceeding essentially *inferentially*, moving from how someone is behaving to what emotion they have, relying on the assumption (in some cases more reasonable than others) that their observable behaviour is likely to have been caused by a specific type of emotion. It is in other words not immediately obvious that any *direct meaning* relation needs to be appealed to to explain the intuitive distinction we draw between direct *expressions* of emotion and so called mere *effects* of these same emotions.

Note that by a '*direct meaning*' relation, I mean essentially one of a kind that might hold, say, between an utterance of 'it is raining' (in the mind of a native speaker) and the proposition that it is raining, in contrast to the kind of *indirect* meaning or 'indicating' relation that might hold between certain signs/clues such as people walking around with umbrellas, and the proposition that it is raining. Put in practical terms, in the latter types of cases, upon seeing the relevant signs/clues, one will generally infer that it is raining *both* on the basis of what one sees *and* on the basis of a further assumption about what generally *causes* or *leads* people to walk around with umbrellas. On the other hand, upon hearing someone say 'it is raining', one will not generally go through any inferential process, in deciding that it is raining, or in deciding that this is what is being suggested. No hidden premise will be appealed to to the effect that when people make noises of the kind "it is raining" they generally intend to convey the idea that it is raining. Rather, one will just directly *hear* their utterance as meaning, and their uttering of it as suggesting, that it is raining.[1]

So, the possibility raised in the paragraph before last in effect amounts to the suggestion that the sense we have that there is some kind of a *direct meaning* relation between having a so called 'guilty' type of look in one's eyes and feeling guilty or between smiling and being happy, is an illusion. In reality, this apparent meaning relation between behaviours we call 'expressions' and their corresponding emotions is no more than an *indirect* 'indicating' relation, just as might be that between twitching and being angry, or between trembling and being afraid, or between people walking around with umbrellas and it raining. The only difference lies in our being more likely to immediately infer (and more justified in doing so) that a person feels guilty or happy upon seeing them give us a 'guilty' type of look or a smile, than to infer that they are angry or afraid upon seeing them twitch or tremble. For a number of obvious reasons though, drawing the distinction between 'expressions' and 'mere effects' of emotions in this way is not entirely plausible. It is not plausible, that is, to think of the relation between behavioural 'expressions' and the types of emotions of which they are expressions, as a purely *indirect* 'indicating' relation.

For one thing, our being justified, and/or our being very quick to infer from someone's behaviour that they have a certain emotion does *not* seem to actually be enough for us to be prepared to count their behaviour as an 'expression' of their

[1] For a more in depth discussion of such different senses of 'meaning' and 'indicating' see (Grice 1957).

emotion. Suppose for example that I know that Mary tends to twitch in a peculiar way when she is angry and *only* when she is angry. Upon seeing her twitch in that way I will therefore tend to very quickly infer that she is angry, and be moreover entirely justified in doing so. Yet, it is not obvious that her twitching in this way is something we would want to call an 'expression' of her anger, rather than just an unusual accompanying *side effect* of it. Of course, one might insist that this is so only because twitching as a result of being angry is not as wide spread an effect of anger as is shouting or having a certain 'angry' type of look in one's eyes. We might therefore just *not yet* have come to see twitching as directly suggestive or evocative of anger in the way that we have come to see certain looks in people's eyes, or specific tones of voice as directly evocative of the same emotion. Having said that, to point this out is to make an essentially *developmental* point. The end result, relevant here, remains the same. Some forms of behaviour seem to have come (be it through observed constant association, cultural conditioning, or as a result of innate hard-wiring) to immediately *evoke*, *mean*, or *designate directly* (ie. not by way of any additional assumption about their causal origin) certain types of emotions or the fact that someone *has* these emotions. The move from perceiving behaviours we refer to as 'expressions' to attributing emotions to people on that basis is, phenomenologically at least, a distinctively *non-inferential* one. We do not refer back to any knowledge of constant associations between bodily movements of specific physical types and specific types of emotions when attributing emotions to others on the basis of observing them do such things as smile, laugh, etc. Rather, we seem to directly *perceive* their bodily movements as suggestive of emotion, or, as we might now say, as '*expressive*' of happiness, amusement, etc.

This becomes all the more obvious when looking back again at cases of emotional pretence and at our attitudes towards them. When watching someone simulate a supposed behavioural 'expression' of emotion (eg. a smile, a frown, etc), it is not just that we are easily fooled into believing that they actually have the corresponding emotion. But, even in cases where we *know* that someone (say, an actor) is only pretending, we *still* tend to often think of their way of moving, tone of voice, etc. as being, as we in fact often say, 'highly *expressive*' or 'emotionally very *suggestive*' and so on. In other words, we seem to often think of particular ways of behaving as *themselves* expressive of specific types of emotions, quite independently of whether a person behaving in these ways actually has the emotions in question.

And, as we have seen, this being so fits well with the phenomenology of the process of actually attributing emotions to others on the basis of observing their behaviour. When someone smiles, laughs or gazes at us in a particular way for instance (whereas not when they twitch or tremble), we often feel that their emotion itself is 'transpiring' or 'coming through' in their behaviour. We often say moreover such things as that a person's anger or resentment can be *seen* in their eyes or *heard* in their voice. People just *look* angry to us or *sound* resentful. It is as if an inner state of theirs (or the *type* of state that they are in) comes to be directly 'written on their face', and so somehow immediately outwardly available to us.

Taking on board these observations about the directly 'expressive' character of our so called 'expressions' of emotion, the notions of 'expression' and of 'expressiveness' can now start to be seen to come apart, yet at the same time to be closely related. 'Expressions' of emotion may in fact now be re-defined thus: for an instance of behaviour to constitute an *expression* of emotion as opposed to an *action* out of it and as opposed to a mere involuntary physical *effect* of it, it must be both directly *caused* by a particular emotion, and intrinsically (ie. non-relationally) of a behavioural type that is directly *evocative* or *expressive* of the type of emotion of which it is a direct effect/manifestation.[1] It is then in virtue of the intrinsically emotionally *expressive* character of people's *expressions* of emotion that we are able to tell what emotions they have directly on the basis of observing this class of their behavioural outbursts. We may thus have here not just a definition of 'expressions' of emotion, but also the beginning of an account of our knowledge of other people's emotions, an account according to which this knowledge is often *direct* and distinctively *non-inferential*, though of course not authoritative. The primary aim of this chapter was however to account for our knowledge of our *own* emotions. So, how might this fit in?

---

[1] Note that this is not to be taken as some kind of 'add-on' theory of expressions, whereby an episode of behaviour counts as an 'expression' of an emotion if and only if two criteria *coincide*, ie. that the behaviour is *caused* by an emotion and that it simultaneously happens to be *expressive* of this same type of emotion. I am inclined to think (in agreement for instance with Wollheim – see Wollheim 1999, ch.1) that emotions are dispositional states, and dispositional states moreover a *constitutive* part of which is the disposition for these states to manifest themselves episodically (whether outwardly or in consciousness) in ways (or in episodes themselves) that tend to be distinctively *expressive* of the types of emotions thereby manifested. Being disposed for example to smile in particular circumstances (ie. to do something expressive of happiness) is both causally and conceptually *constitutive* of being happy – one could not count as being truly happy if one were not disposed in this way. The status of our emotions as such dispositional states, and what exactly is to be understood by this, will be made clearer in the next section (pp.179-180), though it is not part of the main aim of this chapter to provide a worked-out theory of the nature of emotional states – only of the nature of some of their episodic manifestations in behaviour and in consciousness. Of course, to understand the latter, various assumptions made about the former will also have to be sketched out. This will therefore be done briefly further below.

We have seen that we do *not* come to know our own emotions (not all of them at least) through observing our behaviour, whether inferentially or otherwise. Often in fact, in self-ascribing certain emotions we do not seem to look at our behaviour *at all*. Rather, we tend to look out at the *world* for evidence of what emotions we have. So how can any of the above be relevant? Well, emotions do not just find expression in outward behaviour. In what follows, it will be suggested that something directly parallel to what occurs in coming to know *others'* emotions can be shown to occur in coming to know a certain range of our *own* emotions, by reference this time to expressions of emotion in *consciousness*. That is, although we do not read off our own emotions (in certain cases at least) from their emotionally expressive manifestations in our *behaviour*, it will be shown that we read them off nonetheless from their equally emotionally 'expressive' or 'evocative' manifestations in *consciousness*, and in particular from such expressive manifestations (however complex) in our *world*-directed conscious experiences, memories, fantasies, etc. More will of course need to be said about this expressive dimension of our world-directed conscious states if any sense is to be made of our ability to directly read off our own emotions from the way the world appears to us. In particular, the emotional tone taken on by the world of our experiences will have to be shown to be not just emotionally expressive, but to be, from within our own outward looking point of view, recognizably expressive of *ourselves* as having a very specific emotional attitude towards the world.

First though, given the central role that this notion of the emotionally 'expressive' character of our conscious states (or of the world from our point of view) is set to play, the notion itself is worth exploring a bit further on it's own.

What is it for a phenomenon to be 'expressive' of an emotion?

Put in the most general possible terms, we have seen that something can be said to be *expressive* of an emotion essentially if it is somehow *directly suggestive* or *directly evocative* of a particular *type* of emotion – eg. fear, anger, sadness about something, etc. We have also seen more specifically that having this characteristic of being *expressive* of a certain type of emotion is a *constitutive* part of an episode of behaviour's being a direct *expression* of an emotion one now has (as opposed to a mere effect of it). An important difference however has also emerged between 'expressions' of emotion on the one hand and 'expressive' behaviour on the other.

Although we have seen that for something to be an *expression* of emotion it must be also *expressive* of the type of emotion of which it is a direct expression, the reverse dependence does not hold. For a behavioural event to be *expressive* of an emotion it *need not* also be an *expression* of emotion. Being emotionally 'expressive' (unlike being an 'expression' of emotion) is not a property relational upon anyone's now *having* an emotion.

This came out in particular when considering our attitude towards the performances of actors. When praising an actor's performance, we often say, recall, such things as that his or her performance is 'highly expressive' or 'evocative' or 'emotionally very suggestive', etc. Now in so doing, we are clearly *not* saying that their true feelings are transpiring in their behaviour. That is, we are not praising them for letting out their personal emotional baggage in front of an audience or in front of a camera. We are not praising them for bravery (or, bizarrely, for lack of self-control), but rather for skill in the art of expressive movement, suggestive tone of voice, etc. Thus clearly, for us to count an instance of behaviour as '*expressive*' of an emotion, it need not be a genuine *expression* of anyone's actual emotion. The person behaving in the so called 'expressive' way need not have an emotion of the kind being conveyed, nor indeed have any emotion *at all.* Being expressive is a property intrinsic to the behavioural type itself. And in a sense therefore, all expressive behaviour can be said to be *genuinely* expressive, in that it's being expressive is entirely independent from whether or not the person behaving in the relevant way is being sincere.[1]

This independence of emotional expressiveness from being an expression of emotion in fact goes even further. For something to be expressive of an emotion, it need not it seems actually be a piece behaviour at all, nor indeed a state of a *person* at all, whether conscious or otherwise, thereby detaching the phenomenon of emotional expressiveness even more from any necessary association with the idea of being caused by someone's actual emotion. Music for instance, can be thought of as emotionally expressive in virtue of its purely musical properties (eg. rhythm, modality, etc.), properties which it has entirely independently of anyone's mental or

[1] In mentioning this I have in mind the fact that Goldie contrasts what he calls 'genuinely expressive' behaviour with '*non*-genuinely expressive' behaviour. See (Goldie 2000, ch.5). From what we have seen here though, once one disambiguates between speaking of 'expressions' of emotion and speaking of emotionally 'expressive' behaviour, the term of 'non-genuine' can no longer truly be said to apply to instances of the latter. *Expressions* of emotion can be genuine or non-genuine depending on whether or not they actually arose from emotion. *Expressive behaviour* on the other hand remains expressive whether or not the person behaving in that way actually has any emotion at all.

physical states, and indeed perhaps most relevantly, independently of anyone's performance of the piece of music in question. The 'minor' musical modality for example is often thought to be evocative of sadness, and indeed, pieces in so called 'minor keys' tend to be almost universally experienced as sad, or experienced as expressive of sadness. Pieces in 'major' keys on the other hand tend to be experienced as evocative of joy and confidence, while dissonant chords are often suggestive of horror or suspense. Take for example the use of music in films. Specific pieces of music tend to be chosen by filmmakers for specific films very much in accordance with what emotional context or emotional tone they want to set to the film or to particular scenes, that is, in accordance with what kind of emotional spin they want to put on the portrayed events. Horror films are for instance a very obvious context in which music is made use of for such purposes.

An interesting fact about this moreover, is that our perception of the emotional tone set by different types of music does not seem to vary greatly across populations or individuals. It is very unlikely that a particular piece of music, as used in a film, will be misinterpreted as indicating a completely different emotional context from the one intended. Even a sequence of notes as simple as a minor scale is going to be almost universally perceived as evocative of sadness, thus suggesting that our perception of particular musical modalities, rhythms, soft/loud contrasts, etc. as distinctively expressive of specific types of emotions rather than others, is likely to be due more to something constitutionally hard-wired within us, than to any incidental past associations we might have made between specific types of music and particular emotional situations. Put crudely: we do not seem to get the idea that particular pieces of music or sequences of notes indicate sadness or suspense by watching films; rather, specific pieces of music are chosen precisely *because* of the emotional tone that they will set to a scene. Exactly which way around the story goes however for different expressive phenomena (forms of behaviour, types of music, colours, phenomenological distortions of the world, etc.) is of course again an essentially developmental matter, the result of which, relevant here, remains the same. For present purposes, even if the specific expressive emotional tone that a particular type of music sets is the result of learnt associations, it still remains that when actually hearing a piece of music we do not refer back to these past experiences and associations, and infer from this that a certain emotion is being indicated, or that a scene in a film is to be interpreted as a scary one. Rather, we seem to recognize the

relevant musical modalities, rhythms, contrasts, as distinctively expressive of particular types of emotions somewhat *independently* of any past associations we might have made, and indeed often quite independently of having made any associations at all between the particular types of music and specific emotional contexts. More than that, we often distinguish between the type of emotion that a piece of music 'objectively' suggests or is 'objectively' expressive of, and the type of emotion it brings out *in us* given our past experiences. We may for example recognize a particular piece of music as 'objectively' expressive of joy (ie. as expressive of joy in virtue of it's purely intrinsic musical properties), yet still be filled with sadness upon listening to it because it is associated for us with a particular sad event in our lives.

A question that this leaves us with is that of what exactly it is *non-developmentally* about certain specific musical properties, or physical properties, or visual properties of different expressive phenomena (eg. sequences of notes, bodily movements, or, say, the shade of blue), that, from our *personal*-level point of view, right *now*, makes us see it as objectively suggestive of a particular emotion rather than another. Perhaps there is no answer to be given to this. That is, it may just be that from a personal-level point of view, the emotionally expressive character of certain forms of behaviour, certain types of music, certain colours, etc. is just *primitive*, ie. cannot be analysed any further or explained in terms of anything else. Whether or not this is so however will have to be left aside for present purposes. It *will* be suggested though, that when turning to the case of the expressive transformations effected by our emotions onto the *world* of our conscious experiences, there *is* more to be said about what might make these types of transformations distinctively emotional transformations, and indeed *recognizably* emotional transformations from within our own outward looking point of view. That is, it will be suggested that the emotionally expressive character of our conscious experiences, memories, fantasies, etc. *can* be analysed or broken down a bit further, or sufficiently at least to bring to light our ability to look out at this emotionally transformed world, and on that basis alone, self-ascribe the very emotions effecting the transformations.

Before turning to examine more closely this emotionally expressive character of our conscious experiences however, one further phenomenon, and in particular its potential status as an *expressive* phenomenon (like some of those considered above),

remains to be looked into, namely, as promised, that of symbolic or displaced intentional *action out of* emotion.

'Expressive' actions out of emotion

Recall some of the examples of intentional actions put forward at the very beginning as potential cases of direct 'expressions' of emotion. One of these, borrowed from Goldie, was that of tearing the eyes out of a photograph representing someone one is angry at. Another was that of putting pins in a doll resembling someone one dislikes. Such actions, it was suggested *could not* (being by their very nature *intentional* – ie. actions *motivated* by emotions) be at the same time direct *expressions* of emotion (the latter being by their very nature spontaneous and involuntary). But, in light of what has now been said, such actions *could* conceivably turn out to be *expressive* actions out of emotion.

Such diverse phenomena as pieces of music, episodes of smiling performed intentionally by actors, and even colours, it has been suggested can be thought of as directly expressive of specific types of emotions despite clearly not being in any way direct involuntary effects of anyone's emotions. The expressive character of a bodily movement in particular was shown to be entirely independent of whether the person carrying it out is doing so spontaneously and involuntarily or whether they are carrying it out intentionally. Intentional actions such as those considered by Goldie may therefore equally well turn out to be expressive actions despite not being direct expressions of emotion, so long as they are in one way or another of a behavioural type that is *directly suggestive* or *directly evocative* of the type of emotion in question.

Take the case of tearing the eyes out of a photograph representing someone one is angry at. It might be, as indeed Goldie suggests, that there is something primitively intelligible for us about behaving in such ways when angry at someone (ie. attacking them, or by extension, destroying something that symbolizes them). Now, this may in turn result in our seeing such destructive actions as *directly expressive* of anger towards the person represented or as *directly expressive* of a desire to hurt them. It might in fact be that, upon seeing someone behave in such a way, we are immediately inclined to take their acting in this way to *mean* that they *are* angry at the person represented in the photograph. Moreover, and contrary this time to what Goldie actually suggests, such actions, in being expressive in this way

(assuming they are so) will always be so *genuinely* even if the person carrying them out is not actually angry. An action may of course genuinely or non-genuinely *arise* out of emotion (whether motivationally, or spontaneously as in the case of expressions), but, this will have nothing to do with its status as *expressive*. Smiling for instance, just like the sound of a major chord will *always* be expressive of happiness even if carried out intentionally by an actor. And, so it will be also for symbolic expressive actions such as the above.

Thus, despite not being potential 'expressions' of emotion, displaced actions out of emotion of the kind that Goldie considers may well turn out to be *expressive* actions out of emotion. And, in being expressive in this sense, they will always be so genuinely, even if they do not on a particular occasion arise out of emotion. Much more of course could be said about the intrinsically expressive character of such actions. More could also be said to do better justice to what Goldie in particular says about such cases. For present purposes though most of this will have to be left aside. To make a few brief points nonetheless, in light of some of the criticisms we have raised against Goldie, it should be said that his view that symbolic intentional actions out of emotion can also be *expressions* of emotion, is not driven solely by confusion and a failure to distinguish between 'expressions' and 'expressive behaviour'. In actual fact, his view that some intentional actions can also be direct expressions is justified by a particular account he puts forward of what it takes to be an 'expression' of emotion, an account which contrasts somewhat with the one I have been suggesting here – but which is itself, it seems to me, driven to some extent by the above confusion.

Goldie defines 'expressions' essentially as instances of behaviour carried out for no *ulterior* motive, that is, for no reason *other* than say, because one is angry. Certain actions *out of* emotion (ie. actions *motivated* by emotion) *can* therefore on his account be seen to be not just 'expressive' actions but also 'expressions' of emotion, so long as they are not performed as means to any further end. Suppose for example that my anger motivates me to hurt someone. Given various social and psychological taboos however, I end up, as in Goldie's example, tearing the eyes out of a photograph representing the person I am angry at rather than attacking the person herself. Now, in doing this, I am in a sense still acting *out of* the motivational desire which is intrinsic to my anger, that is, out of my desire to hurt them. Yet, since I choose to act out on my desire (or on my anger) merely *symbolically* on a photograph

rather than on the real person, I cannot be said to have acted *in order* to achieve the desired end of physically harming the person represented, since I knew fully well that damaging the photograph would not achieve this. Thus, although my action is intentional, ie. voluntary and *motivated* by my anger and my desire to hurt a particular person, it is not carried out as a *means* to achieve the *end* or *goal* of hurting the person I am angry at. On Goldie's account, this action can thus come out as being a possible 'expression' of emotion. Intuitively though, this does not seem to be enough to make my emotionally motivated action a direct expression of emotion.

To bring out this intuition, consider the following analogy. Suppose I have a strong desire for ice cream, yet go to my refrigerator and discover that there is none left. Suppose I then express some disappointment and have a piece of fruit instead. In eating this piece of fruit I am performing a displaced action out of my desire for ice cream, displaced this time not because of any psychological or social taboo, but simply because doing the real thing is not possible. So, doing something close to the real thing, ie. something similar in form to what would be involved in satisfying my desire seems somehow better than doing nothing. It is far from clear though that my eating the piece of fruit here is something we would intuitively want to call a manifestation or expression of *desire* – in the way that, say, a look of longing on my face directed towards an empty box of ice cream might be. In intentionally eating the piece of fruit I am not (nor is it my *aim* to be) essentially 'airing' my desire or letting it 'come out' in the open. Rather, I seem to be just acting out in a displaced way on my desire for ice cream, in an attempt to reach some (albeit partial) sense of *fulfilment* of my desire (note, not of *release*), through a displaced enactment of what it would take to satisfy it. No direct outburst of emotion or desire need actually be involved in the process. In fact in this example the only emotion seemingly being expressed at any stage is disappointment and perhaps resignation, not desire (ie. the emotion out of which it arose).

Similarly, I would now say, with the case of tearing the eyes out of a photograph of someone I am angry at. My destroying a representation of someone I am angry at, even if done *because* I am angry at this person or *because* I want to hurt them, is not necessarily in itself a direct *releasing* of anger but a mere case of acting out (albeit symbolically) on my anger at the person represented in the photograph. I could be doing this without letting out any anger at all, that is, carrying it out perhaps merely as a symbolic gesture to give myself some sense of closure on my desire. Or, I

could actually be on the verge of letting go but purposely *suppressing* my emotional outbursts – perhaps because I am afraid of how out of control I might get, or perhaps simply because I am the kind of person who has difficulties in letting out my emotions. Of course, on other occasions, my doing something like destroying a representation of someone I am angry at, might be something I do on purpose to provide myself with an *opportunity* or appropriate context in which to let my anger spontaneously come out. I might in fact find it easier to let out my anger in the process of symbolically acting out on it, than in the process of doing something less relevant (eg. walking down the street). However, *allowing* an emotion to come out, or doing something *in order* to allow an emotion to come out, is not *itself* a releasing of emotion, as we saw earlier when discussing the intentional control we have over our emotional outbursts.[1] My anger may well end up being released simply in the particular aggressive *manner* in which I hack at the photograph, but it *will not* be released in virtue of my *intentionally* carrying out this displaced action out of emotion. Perhaps one could argue that in some cases doing things like destroying a representation of someone one dislikes is something one does entirely spontaneously rather than for the *reason* that one is angry or for the *reason* that one wishes to hurt them – ie. one might just be *lashing out* at the photograph (just as one might at the real person) in a fit of rage. If so, that is, if the example in question is genuinely one of spontaneous, non-intentional, non reason-grounded behaviour from emotion, I would be more than prepared to call it a direct 'expression' of emotion – or at least a borderline case. To the extent however that we are thinking of the action (as Goldie is) as an *intentional* action *motivated* by an emotion, it cannot it seems also be at the same time a *direct expression* of emotion. One's emotion might be let out spontaneously in the process of carrying out such actions, but, the intentionally carrying out of them itself is not a direct manifestation of emotion. The point seems to remain that symbolic or displaced *intentional* actions out of emotion, when considered purely as such, are *not* (and are incompatible with being) themselves *direct expressions* of emotion.

This issue could no doubt be debated further, and various further ambiguous/borderline cases brought in. I will leave it aside though at this point, and, on grounds of greater plausibility, continue to take expressions of emotion to be

[1] See p.158 above

essentially direct, spontaneous, involuntary, impulses from emotion, to be *contrasted* with actions *motivated* by emotions, the latter being performed *because* one feels a certain way, as opposed to for no reason at all like the former. One does not smile because one is *motivated* to do so by happiness; one just 'does it'. And, if one *does* smile because one is happy (ie. for the *reason* that one is happy) one is *not* in virtue of that fact alone also venting one's happiness in the process. And, so it will also be if one does anything else intentionally, no matter how much *what* one does (eg. smiling, or tearing the eyes out of a photograph) might be directly *expressive* of a specific type of emotion.

The important point to bear in mind here is that even on the present understanding of expressions (ie. as strictly involuntary forms of behaviour), symbolic intentional actions out of emotions can, in agreement with Goldie, be thought of as actions *expressive* of emotion. Tearing the eyes out of a photograph representing someone one is angry at does in fact seem to be directly suggestive of anger, or of desire to hurt the person represented, or as Goldie puts it, expressive of a 'wish' to hurt them. Recall, an action's being *expressive* of a type of emotion requires only that there be some kind of direct *conceptual* link between the course of action taken (damaging a representation of someone) and the type of emotion in question (anger at the person represented, or desire or wish to hurt the person represented). Standing in a direct *conceptual* relation to a *type* of emotion however or being *expressive* of a certain type of emotion does not in itself turn that which has this expressive character, automatically into a direct *expression* of emotion. The latter depends essentially on how one's behaviour arises from one's emotion, ie. spontaneously and involuntarily or in a reason-based, motivated manner.

Moving away now from actions out of emotion, and from expressions of emotion in behaviour, as well as from the general phenomenon of emotional *expressiveness* in behaviour, music, colours, etc., the time has come to focus more narrowly on the case of expressions of emotion in *consciousness*, and in particular, armed with the above conceptual framework and distinctions, on the emotionally expressive ways in which our emotions, in manifesting themselves in consciousness, come to transform the *world* from our perspective, this world which is to provide us with direct evidence of ourselves as having specific emotions.

7.2 Manifestations of emotion in consciousness

An obvious first way in which our emotions seem to manifest themselves in consciousness is in *sensation.* When we are afraid, angry, happy, in love, etc. this can often result in our coming to have a wide range of sensations, such as feeling flushed, unable to breathe, having chest pains, feeling faint, or, in less extreme cases, in our just feeling physiologically aroused in some vague and generalised way. This is in fact such a common part of our everyday experience of emotion that it has been thought by some to lie at the very heart of what it *is* to have an emotion. According to William James for instance, emotions do not just sometimes *manifest* themselves in sensations, but they essentially *are* sensations, specifically sensations caused by the perception of visceral disturbance or other physiological change.[1] For a number of reasons though frequently raised in the literature, this does not seem to be the best way of thinking of the nature of emotions, nor indeed to be the best way of thinking of how our emotions in many instances even just manifest themselves in consciousness.

For one thing, as pointed out by Cannon,[2] specific visceral disturbances do not correlate accurately with specific types of emotions. One same state of arousal or physiological disturbance may on different occasions be associated with a number of different emotions, and conversely, different states of arousal may be associated at different times with the *same* emotions. In fact, physiological arousal states are often not associated with (nor even accompanied by) any emotional states *at all.* Feeling one's heart beating very fast for instance, may be associated equally well on different occasions with fear, love, surprise, or with just having done some intensive physical exercise. Similarly, feeling flushed may be associated with being embarrassed or angry, or on the other hand with just being hot. And, perhaps most strikingly, having chest pains and being unable to breathe can be associated on occasion with extreme anxiety, yet on other occasions with one's just having a straightforward heart attack. In other words, an account of the nature of emotions along James's lines that focuses primarily on the 'sensation' and 'visceral' aspects of our emotions will ultimately be unsatisfactory (a) in being unable to adequately distinguish between different types of emotions, and (b) in being also unable to adequately draw the line between emotions and non-emotional sensations caused by the same visceral disturbances.

---

[1] See (James 1884)

[2] See (Cannon 1929)

Beyond Cannon's objections, a 'sensation' account of the nature of emotions also faces a number of other problems. First, insofar as emotions are on this view essentially *conscious* states, there will be little room made for the existence of *non-conscious* or *repressed* emotions, say, repressed feelings of guilt, or resentment, or desire, that most of us feel do exist and can be identified through observing people's patterns of behaviour. This is a problem that will be shared in fact by any account according to which emotions are essentially *conscious episodes*, whether sensations, judgements or perceptual experiences. Another problem, more specific this time to a 'sensation' account of emotions, is that it will it seems leave out the whole *cognitive* dimension of our emotions. Emotions are states we generally think of as potentially directed towards a wide range of external objects, events, states of affairs, etc. and not just towards our bodies and our bodily states. One can feel resentment towards another person or happy about some external state of affairs, or afraid that some outside event might occur, and so on. Might there be ways around these problems?

Perhaps one could try to accommodate the cognitive element by incorporating it somehow on the side of the *causes* of emotions, as done sometimes for instance by 'appraisal' theorists in the recent psychological literature.[1] That is, one could perhaps say that emotions *themselves* are merely bodily sensations, though they are sensations specifically caused by particular world-directed cognitive states (eg. evaluative beliefs or judgements), the latter being relevant to their individuation. Taking this kind of line though would leave in place the problematic idea that in *having* an emotion, at the conscious level, one's attention is essentially directed towards one's visceral states. That is, taking this line would be to hold on to the view that the conscious content of our emotional experiences is still the same as that of sensations.[2] Yet, when actually considering the conscious content of our emotional experiences, it does not seem that in most cases of feeling angry or happy or in love the focus of our conscious attention is primarily on our body or bodily states. Rather, when afraid of something or happy about something, or in love with someone, our attention tends to be focused (if on anything at all) out on the *world*, and in particular on the *objects* of our fear, happiness, love, etc. In being afraid of a rat for instance I will tend to be more focused

[1] Appraisal theories will be discussed in further detail in the next chapter.

[2] I am contrasting here the *conscious* content of an emotional experience with the *intentional* content of the emotion itself – the latter of which may be cashed out in a number of ways, and most relevantly *need not* actually correspond (most obviously in the case of unconscious states) to anything one actually has *present to mind* in having this state at any particular time.

on the *rat* than on the fact that I am trembling. It is only in very unusual cases that our attention, in having an emotion, will be turned in the first instance towards our bodily states. An example might be that of having a repressed feeling of anxiety, which may result in one's having an anxiety attack, yet which might, in being repressed, manifest itself in consciousness merely as awareness of chest pains and inability to breathe, often in fact to the point of being mistaken for a purely physiological state – ie. a heart attack. The conscious phenomenology of emotion is however *not* generally the typical phenomenology of awareness of visceral change. Our feelings of guilt tend to manifest themselves in our being completely absorbed by the wrong actions we have committed, rather than with our current bodily states; our feelings of fear tend to manifest themselves in our having in mind the objects of our fear rather than the fact that we are trembling or can't breathe; and, even feelings of so called *free-floating* anxiety in most cases manifest themselves in consciousness primarily in *world*-directed concerns and *world*-directed perceptual and cognitive focus, rather than in self-focused feelings of physiological discomfort.

This brings us back to a point mentioned earlier. As noted in the neuroscientific literature, emotions do not just give rise to particular arousal symptoms and sensations, but they manifest themselves to a great extent also in directing our attention – ie. in determining the salience of certain things rather than others in the environment, in making certain past events stand out in our memory, in leading us to imagine certain scenarios rather than others, and in making us envisage particular possibilities and possible courses of action when deliberating about what to do.[1] Thus, not only are emotions *themselves* not mere sensations, but mere sensations do not seem to be the only mental phenomena in which emotions *manifest* themselves

[1] When these points are made in the neuroscientific (or indeed philosophical) literature, they are usually made within the context of concerns about the *function* of our emotions in our mental lives and specifically in our *decision-making* processes. According to Damasio (1994) and De Sousa (1987) emotions play the *crucial* and *in principle indispensable* role of bringing to an end lengthy deliberative processes, and of sometimes *supplanting* such deliberative processes altogether in decision-making circumstances where time is of the essence. The idea is, first of all, that without emotions automatically directing the focus of our attention to certain things rather than others and to particular courses of action rather than others, we would be unable in some cases to *ever* stop deliberating and come to a decision. Then, the idea is (according to Damasio in particular) that not only do emotions save us from decision-making deadlocks, but the possible courses of action that our emotions direct our attention to tend to be the very ones we *should* be attending to, that is, the most *rational* or *adaptive* courses of action for us to take. I have doubts about the idea of our emotions being *in principle* indispensable in getting us out of decision-making deadlocks (eg. could we not use methods such as tossing a coin instead?). I also have doubts about the idea of emotions always drawing our attention in practice (let alone in principle) to what is most *relevant* or most *adaptive* (eg. does a phobic fear of feathers make us focus on what we *should* be focusing on?). Nonetheless, it does seem true to say that, for better or for worse, emotions *do* determine the focus of our attention to a great extent, be it in perception, memory, or thought. They do *not* therefore just manifest themselves in self-focused sensations, but also often in giving rise to specific *world*-directed conscious episodes and in presenting certain courses of action to us as *the* ones to take.

in consciousness. This is of course not to deny that sensations are amongst the many ways in which our emotions might manifest themselves in consciousness, sometimes even on their own, as in the repressed anxiety attack case mentioned a paragraph back. But, it seems that they cannot be, and indeed are not (purely as bodily sensations) the *only* kinds of direct effects that our emotions have on our mental lives, nor therefore the only contexts in which our emotions might come to find expression in consciousness. That this is so is fortunate moreover in that the 'sensation' aspect of our emotional experiences is not an aspect of the ways in which our emotions find expression in consciousness that could explain our ability to *know* our own emotions directly on the basis of looking out at the *world.* In fact, even if we did base our knowledge of some of our own emotions on some form of inference from (or indeed direct perception of) our arousal symptoms and sensations, this way of knowing our own emotions, being based essentially on data explicitly about *ourselves* and our *states*, would not explain how it is that we actually can, as seen in chapter 6, also sometimes come to know our own emotions on the basis of evidence that is in the first instance about the *world.*

But, one might now ask, can the newly introduced fact that emotions manifest themselves in conscious episodes other than sensations, ones directed towards the world, do the trick? It is not immediately obvious that it can. After all, the mere fact that I am finding my attention continually drawn to certain things rather than others may well give me (if even that) a clue to the fact that I have *some* emotion or other towards that which I am continually drawn to, but, it will not in itself tell me *what* emotion I have. Moreover, even to the extent that it will be able to tell me this, the road to knowledge will be essentially inferential and based again on evidence explicitly about *myself* and *my states* – evidence about what I am thinking about, what I am looking at, where my attention is continually drawn to etc. – rather than being based on evidence in the first instance about how the *world* is from my perspective. The solution to our puzzle therefore remains still to be found.

Before moving on though, it should be stressed that what has now clearly become the issue, and what we need to explore in order to solve our puzzle, is not what emotions themselves *are* but the different ways in which our emotions *manifest* themselves or find *expression* in consciousness, and this for two reasons. On the one hand, it has been denied above that emotions can themselves be conscious episodes of

any kind.[1] Yet, on the other hand, what we clearly need to identify is something about the *conscious phenomenology* of looking out at the world in having emotions, since this is supposed to be the starting point of our knowledge of our own emotions. What we therefore need to examine, if not our emotions themselves, is nonetheless their contribution to our experiences of the world. But now, if this is what we are looking for, ie. if what we are looking for is to uncover a specific way in which our emotions are *expressed* or *transpire* in our world-directed conscious experiences, we might first want to know what kind of states the supposedly 'expressed' or 'transpiring' emotions themselves are. In other words, to make a brief aside, what *are* emotions, if not conscious episodes of some form? In what follows, and in light of the criticisms raised here of the views that they might be some form of conscious episodes (eg. sensations or judgements), the following broad assumptions will be made.

I am taking it, and will continue to do so in what follows, that emotions are in the first instance *dispositional* states, ie. dispositions to behave in various emotion-specific ways and to have a wide range of other conscious episodes and non-conscious states that are characteristic of the type of emotion in question. Importantly though, in saying that emotions are essentially dispositional states, and in particular, in assuming that a tendency to have a wide range of specific conscious and behavioural episodes is conceptually and/or causally *constitutive* of having these dispositional states, I do *not* mean to say (as Ryle would for instance)[2] that to have emotions just *is* to have particular patterns or sequences of states. On the contrary, I am assuming emotions to be, put in Wollheim's terms, somehow 'psychologically *real*',[3] that is, to be internal states of some kind *in virtue of which* we are disposed to behave, think, experience the world, etc. along certain patterns. My smiling for instance is not just constitutive of my being happy, but my being happy is also what *caused* me to smile. My emotions can be appealed to to *explain* why I am smiling. On this understanding of emotions in other words, it still makes sense to speak of our

[1] In this chapter it has been argued that emotions cannot be *sensations*, and in chapter 6 arguments were put forward against cognitivist theories of emotions according to which emotions are particular forms of evaluative *judgements* or (along similar cognitivist lines) *perceptual experiences* of the world as having evaluative properties such as of desirability, blameworthiness, etc.

[2] See (Ryle 1966). Ryle of course speaks only of *behavioural* dispositions, but the principle is the same.

[3] See (Wollheim 1999, ch.1)

emotions as *causing* their expressive manifestations,[1] as *motivating* (ie. standing as *reasons* for) our actions out of them, as being *reflected* or *mirrored* in certain complex experiences or transformations of the world from our perspective, and indeed as themselves *transforming* the world of our experience in ways that *reflect* them (often in all their structure and complexity – eg. by transforming different aspects of the world in different ways). And, all of this can be said while still thinking of these various states as causally and/or conceptually *constitutive* of having the emotions in question. The fact that certain episodes are constitutively linked with having an emotion does not, that is, mean that they must in some sense just be *equated* with the emotion – quite on the contrary if we are speaking of the constitutive link as a *causally* or *conceptually* or *rationally* constitutive one, rather than as one of *identity* or *constitution.* So, to sum up, emotions will be assumed to be essentially states *in virtue of which* we are disposed to behave in various emotionally expressive ways, disposed to attend (be it in memory, thought, perception, imagination, etc.) to specific objects/ events/ states of affairs relevant to our concerns, and ultimately, we will see, to have our perspective on the world 'transformed' or 'coloured' in various emotion-specific (ie. emotionally *expressive*) ways.

With this conception of emotions in mind, we can now return to our main task; that of uncovering a way in which our emotions might manifest themselves in consciousness, at the world-directed level, such that knowledge of our own emotions based directly on looking out at the world might be possible. So far, we have seen that emotions can manifest themselves in behaviour, in sensations, and in giving rise to a wide range of episodic conscious states such as distinctively focused perceptual experiences, episodes of recollection, imagination, particular fantasies, individual thoughts (perhaps recurrent thoughts), obsessive dreams, and so on. None of this however seems to help explain how it is that we are able to come to know our own emotions by just looking out at the *world* – ie. *without* having to consider in the first instance any facts about ourselves and our mental states. If our puzzle about self-knowledge is to be solved, our emotions will have to be found to manifest themselves somehow not just in our coming to *have* world-directed states (ie. in our coming to *attend* to the world, in perception, memory, imagination, etc.), but also essentially in

[1] The causal interaction between emotions and expressive conscious episodes can of course also go the other way around. A sudden emotional conscious episode might establish a lasting emotional disposition, which may then in turn come to manifest itself in various future emotionally expressive episodes.

altering the way the *world itself* comes to appear to us from our conscious point of view in having these states. And, something along these lines was in fact already suggested in the first section of this chapter.

In discussing expressions of emotion in *behaviour*, it was noted that emotions do not just manifest themselves in distinctive episodes of behaviour but also often (and sometimes exclusively) in the *manner* in which the behaviour is carried out (eg. loudly, aggressively, very fast, etc). Similarly, it was then anticipated, it would also be for the case of manifestations of emotion in *consciousness*.[1] That is, it was already suggested that in a somewhat parallel way, when it comes to expressions of emotions in consciousness, our emotions do not just manifest themselves in our coming to have particular conscious episodes that we might not have otherwise had, or in our attention being drawn to specific aspects of the world rather than others, but also to a large extent in the *way* in which the world towards which our attention is directed in these conscious episodes is itself experienced/ remembered/ imagined from our point of view. The example used was that of a remembered event, which, it was suggested, would figure in one's memory experiences very differently according to whether one felt embarrassed about it or proud of it, etc. What should we make of this claim, ie. of the claim that emotions manifest themselves, amongst other things in the way the *world* appears to us from our conscious point of view?

Phenomenologically, to start with, it rings true. Not only does it seem true phenomenologically (though this of course needs to be articulated further), but it also seems to fit well with many of the things we actually say when *describing* our emotional experiences. As noted in some of the recent psychological literature by authors such as Frijda and Lambie & Marcel, when trying to get someone to understand exactly how we feel, we often describe aspects of how the *world* seems to us, or of how the world has *come* to seem to us (eg. gloomy, alienating, bright, open, welcoming, etc.) rather than pointing to anything explicitly about ourselves and our *mental states* (eg. our feeling down, unhappy, overwhelmed, elated, etc.).[2] Often in fact, without even noticing, we find ourselves describing how we feel in *mixed* terms, that is, mixing descriptions of the world as it seems to us with more reflective claims explicitly about ourselves as feeling this or that way. But now, one might ask, how

[1] See p.167 above

[2] See (Frijda 1989); (Lambie & Marcel 2000)

can describing the way the *world* seems to us possibly be an appropriate way of answering a question about how we *feel*?

Whether an emotion is reported more in one way or the other tends to vary with a number of factors such as the context of the report, one's attentional habits (ie. whether one tends to attend more to oneself and to one's mental states or to the world), what type of emotion one is describing – eg. anger apparently tends to be reported in more world-oriented terms, whereas depression is more commonly reported with greater reflective mention of oneself as feeling this or that way. Having said that, it still remains that both types of reports are often produced in answer to the same question, ie. in answer to a question essentially about one's *emotions*. It must therefore be that somehow describing how the world seems to us *can* be informative about how we feel. To put things differently, if we *do* instinctively find that describing the way the world is for us phenomenologically can constitute a direct and appropriate way of answering a question about our emotions, it must be that these descriptions actually are (as are more explicit second-order statements) somehow *directly expressive* of the emotions we are trying to get across. It is only the *way* in which we are getting the message across perhaps that is different – in the one case by *stating* what emotions we have, and in the other by getting our audience to take on our point of view and thereby 'see' how we feel, where 'seeing' how we feel involves reading off our emotions from certain expressive ways the *world* appears to us from our point of view. In fact, if we assume (as it is plausible to do so) that the descriptions of the world we produce in response to questions about how we feel do have some basis in phenomenology (ie. if we assume that they do not just designate specific emotions as a matter of linguistic convention), it must be that, not just the descriptions of the world, but the very ways the world appears to us that we are trying to get our audience to visualise, are (like certain pieces of music, or colours, or forms of behaviour) themselves *directly expressive* of the emotions in question, from within our own outward-looking point of view, a point of view which we can invite others to take on.

That the world is phenomenologically transformed in such ways in emotion experience is by no means unheard of in the literature and is perhaps most vividly described by Sartre.[1] He describes for instance a case of suddenly seeing a face

[1] See (Sartre 1962)

behind a window, which, in one's being terrified by it, appears to one not as a face behind a window, and so as a face at a safe distance, having to get through a locked door to get to one, etc. but as an immediately and primitively *terrifying* face, *framed* by the window (not kept away by it), immediately *acting* upon one at a distance (without requiring any *means* to do so), and so on. The world in Sartre's example in other words ceases to be experienced as a utilizable and deterministic world, and comes to be seized instead as a 'horrible' 'non-utilizable *whole*'. In Sartre's phrase 'the world of the utilizable vanishes abruptly and the world of magic appears in its place'[1] – a description of the world-directed phenomenology of fear that does indeed seem quite reminiscent of the experience of sudden horror. There may of course be aspects of the world-directed phenomenology of emotion experience that Sartre describes in this and other examples that we may identify with more than with others. Nonetheless, whatever is to be made of the specifics of Sartre's position (which will be discussed in further detail below), the simple idea that our perspective on the world, or the world from our perspective, *is* somehow phenomenologically 'transformed' or 'coloured' (or in pathological cases we might say 'distorted') in emotion experience does seem plausible, and is indeed not a view held solely by Sartre.[2]

Wollheim for instance also states such a view quite explicitly. In the context of distinguishing between beliefs, desires and emotions, he writes: 'If belief maps the world, and desire targets it, emotion tints or colours it'.[3] He too in other words seems to endorse (though he does not spend as much time articulating it – his interests lying elsewhere) some form of the view that our emotions affect not just our behaviour, our sensory states, and our attentional focus, but also, and perhaps *primarily*, the *world* as it appears to us.

Most recently, in the psychological literature Lambie & Marcel also put forward the view that there is a distinctive *world*-directed phenomenology of emotion

---

[1] (Sartre 1962, p.90)

[2] Strictly speaking, according to Sartre emotions do not colour or transform the world of our experiences – they *are* transformations of the world. Emotions, for Sartre, are essentially 'ways of apprehending the world', and in particular, they are apprehensions of the world under the mode of the 'magical' (ie. the non-deterministic, the non-utilizable, the absolute, etc). For present purposes though (since we have denied that emotions are themselves conscious episodes of any kind), Sartre's view can be taken simply as a proposal about the phenomenology of emotion *experiences*, without our having to assume that this is all there is to emotion.

[3] (Wollheim 1999, ch.1 p15)

experience in addition to the *self*-focused phenomenology of emotion experience more commonly noted in the rest of the psychological literature. Lambie & Marcel go even further than just advocating this view and describing some of the distinctive phenomenological transformations undergone by the world. They make the further claim that these world-directed emotion experiences and our *self*-directed emotion experiences are often (as suggested to an extent by the indifference with which we report our emotions in world-focused and self-focused terms) just two sides of the same coin. At the first-order non-reflective level, in having an emotion, our attention tends to be primarily turned towards an emotionally transformed *world,* which we can describe as such. But, by a simple shift in our attention we can, instead of attending to this emotionally transformed world, attend to ourselves as feeling the corresponding emotion towards it. For example, we can shift from focusing on the *impellingness* of an object to ourselves as feeling an *urge* to get it – the two types of experiences being essentially attentional counterparts of each other.[1] In such suggestions we may already see emerge a possible solution to our puzzle about emotional self-knowledge, that is, a possible account of how awareness of ourselves as feeling a certain way towards the world might already be implicitly or 'pre-reflectively' present also in our awareness of the world. For this insight to generalise though, what holds for the case of desire for an object will have to be shown to hold also for more complex emotions and to occur in correspondingly more complex transformations of the world. In any case, we can say that there clearly are elements already in the literature on emotion that point towards a solution to our puzzle – even though they are not put forward for that specific purpose.

Despite these sometimes insightful descriptions however of the phenomenology of emotion experience to be found in the literature (which we will return to in detail later), neither Sartre nor Wollheim nor indeed Lambie and Marcel (ie. the prime advocators so far considered of the view that the world is phenomenologically coloured in emotion experience) actually seem to provide any direct account of what it is, in general terms, for the world to be 'emotionally

[1] Regarding *felt urges* to do something, it has been argued by some, Frijda in particular (See Frijda 1989), that felt action urges (at the conscious level) or *action readiness* (more generally) is a crucial component of emotion and of emotion experience. Frijda sometimes goes so far as to take emotions to just *be* states of action readiness. The insight contributed by Lambie & Marcel however is that of pointing out that this action readiness does not just manifest itself in consciousness as a *felt urge* to act, but also crucially in the *world*-directed phenomenology of emotion experience, ie. as an experienced *impellingness* or *luring* power of the *object.* A similar idea is present also in Sartre, when he speaks of the world being immediately experienced as *requiring* things of us or as making *demands* upon us. (Sartre 1962)

coloured', that is, in particular, what it is for the world to be coloured in distinctively *emotional* tones as opposed to just being transformed in *some* way or another. They all seem to assume (or their silence on the matter seems to suggest) that a colouring of the world is distinctively *emotional* simply in virtue of the fact of being *caused* by an emotion, or in virtue of the fact of being of a kind *usually* associated, as a matter of empirical fact, with the having of a certain type of emotion. Sartre for instance would most likely say that for a transformation of the world to be distinctively *emotional*, it must just be a transformation of the kind he describes as actually occurring in emotion, ie. a transformation into the 'magical', where the world comes to be perceived as a 'non-utilizable whole', as *acting* upon one directly at a distance, as potentially making *demands* upon one, etc. This kind of answer however still leaves us without a general understanding of why *these* kinds of transformations should count as *emotional* transformations while other transformations also associated with or caused (perhaps even systematically) by the same emotions might not do so, such as some emotionally-neutral perceptual hallucinations for instance.

Sartre, as indeed Wollheim and Lambie & Marcel do not explicitly address this question, perhaps primarily in that their interests lie elsewhere. Sartre is interested essentially in reaching an account of why the world *comes to be* transformed in the way that it does in emotion, suggesting ultimately that it is *us* who (somehow intentionally, even if not fully consciously) transform the world into the magical as a strategy for dealing with a world experienced as difficult. In being afraid for instance of a ferocious animal coming towards us, he argues that we suddenly transform the world into a 'magical' world, so as to thereby justify a would-be 'magical' solution to our problem – eg. fainting or closing our eyes as a way of annihilating the oncoming animal and thereby getting ourselves out of this otherwise inescapable situation. Wollheim in turn, in a slightly different way, is also interested essentially in the *origin* of the emotional colourings taken on by the world in emotion, rather than in providing an in principle analysis of what it *is* for the world to be emotionally coloured in the first place. His aim though is primarily to identify a *non-intentional* psychological origin of these emotional colourings in desire – specifically in the satisfaction or frustration of a preceding desire. Lambie & Marcel in turn spend much time first of all describing the various ways in which our emotions *actually* transform the world (in light of, amongst other things, the wide range of available data about how we describe the way the world appears to us in emotion experience) but they leave, like Sartre and

Wollheim, essentially unanswered the question of what makes the transformations of the kind that actually occur in emotion experience distinctively *emotional* transformations, rather than *non*-emotional, say, purely visual transformations caused by these same emotions. Let us therefore take on this question ourselves.

What is it, in general terms, for an emotion to give a specifically *emotional* tone to the world from our perspective? What is it in fact for the world to be emotionally coloured whether as a result of emotion or as a result of anything else (eg. a psychoactive substance, a cognitive evaluation, or a direct trigger from the environment)? And, in particular, given the latter question, can it actually be said that there is nothing more to the world's being coloured in a specifically emotional tone, than for the tone or colour taken on by the world to be directly *caused* by an emotion one now has – thus implying that if qualitatively the same phenomenological transformation were *not* caused by a prior emotion, it would not be an emotional transformation of the world?

Intuitively, merely *causing* a transformation or being in some other way associated with an emotion at the physical level (as seen for the case of behaviour) does not seem to be enough, nor indeed necessary, to make the transformation or colouring a distinctively emotional one. First, certain episodes of 'emotional' transformations of the world (such as those described by Sartre for instance) may actually occur *prior* to one's having any established emotional disposition that might have been causally responsible for it. One's fear in Sartre's example might have kicked in *with* the transformation of the world into the magical, at the sight of the face behind the window, rather than having preceded it.

Next, and conversely, some emotions might, amongst other things, actually cause (or be in some other direct physical way associated with) phenomenological transformations of the world that we would *not* think of as specifically 'emotional'. To see this, consider a case of a person Mary for whom, let us suppose, things start to appear blurry whenever she feels anxious. Anxiety in Mary in other words tends to result in the appearance of the world undergoing a very specific phenomenological transformation from her perspective. And yet, despite this systematic correlation between the hazy appearance of the world from her perspective and anxiety in her, it is far from clear that such a transformation (even though by hypothesis *caused* by her anxiety) *is* one that we would intuitively want to call an *emotional* transformation of

the world or an *expression* (as opposed to a mere 'manifestation' or mere 'effect') of her anxiety in consciousness. So, what might be missing?

What seems to be missing is the same as what was missing from non-emotional manifestations of emotions in *behaviour* (eg. twitching in some unusual way as a result of being angry). That is, what seems to be missing is some form of direct *conceptual* link or direct *meaning* relation between the way one's emotion is manifesting itself (ie. in this case in making the world appear out of focus), and the type of emotion thereby manifesting itself in the transformation of the world (ie. in this case anxiety). There does not, that is, seem to be any direct meaning relation between the world's appearing out of focus and one's being anxious, just as there did not seem to be any such direct meaning relation between twitching in some unusual way and being angry. On the other hand, there *does* seem to be some form of direct meaning relation between a face's appearing in the kind of way described by Sartre in his example and the feeling of sudden horror, much in the way that there was seen to be between smiling and being happy or between having an 'angry' type of look in one's eyes and being angry. Put differently, experiencing the world in the sort of way described by Sartre (to stick to the same examples) in feeling sudden horror does seem to be somehow directly expressive or evocative or suggestive of the feeling of sudden horror, whereas experiencing the world as blurry does not in a any obvious way seem to be directly expressive or directly evocative of anxiety.

With the conceptual framework and distinctions (eg. between expressions and expressiveness, between expressions and mere manifestations, etc.) set up in the first part of this chapter, we may now have a way of articulating in less metaphorical terms (than by speaking of 'emotional colourings') what exactly it is for the world to take on a distinctively 'emotional tone' from our perspective. For the world to take on a distinctively emotional colour is, we might now say, essentially for it to be phenomenologically transformed in a way that is *directly expressive* or *directly suggestive* or *directly evocative* of a specific type of emotion. This may occur as a result either of an underlying emotion we already had from before (ie. in which case the transformation would constitute a case of direct *expression* (ie. of expressive manifestation) of our emotion in consciousness), or, as a result of something else, such as an external trigger (the sudden appearance of someone at the window), a cognitive evaluation (our having come to the conclusion that all is doomed), or perhaps simply as a result of our having ingested some psychoactive substance (eg. a

depressant, too much alcohol, etc.). Fear, joy, depression, etc. may in other words arise either *with* a sudden emotional transformation of the world from one's point of view, or, the emotional transformation of the world might be a manifestation of an underlying emotion one already had, which might have now come to be expressed in this way the world appears to one. Either way, in 'emotionally coloured' experience, the *world* appears in a way that is *directly expressive* of how we feel, thus beginning to shed some light on how we might be able to read off specific types of emotions directly from the way the *world* seems to us. Having said that, a crucial problem still remains.

The world's appearing to us in a way that is directly expressive of a specific type of emotion is not in itself enough to tell us, from our strictly outward looking point of view, and without our having to bring in any further assumptions about what is usually associated with the world's being phenomenologically transformed in this way, that *we* ourselves (or indeed anyone else) actually have the type of emotion of which the appearance of the world is directly expressive. To see this problem more clearly, consider the following imaginary scenario.

Suppose for the sake of argument that sadness in me tends to result quite literally in my coming to experience the world as tinted in a shade of blue (ie. as if seen through a blue filter). Suppose also that we agree that the colour blue is a colour directly expressive of sadness, in the way that red is of anger, or in the way that certain musical modalities are suggestive of joy and optimism. What follows then is that when I am sad I tend in effect to start seeing the world in a shade that is directly expressive of sadness, in fact in a shade that I directly *recognize* as expressive of sadness from within my own outward looking point of view. Yet clearly, the world's appearing blue to me *does not* in itself seem to say anything directly about whether I myself am actually sad. All that is implied by the blue appearance of the world from my perspective is that the world looks sad to me or has taken on a form that is directly expressive or evocative of sadness from within my point of view. This neither explains how, nor shows that, the appearance of the world from my perspective contains implicit in it reference to *myself* as being sad. In fact, even if it were true that, as a matter of empirical fact, I could not be seeing the world as tinted in blue unless I was actually sad, this would still not make the world's appearing blue to me constitute a possible *direct* ground for self-ascribing the emotion. At the very least, in order to make this self-ascription, I would have to make the additional assumption that the

world appears blue (or in any other manner directly expressive of sadness) to one only when one is oneself sad, ie. the assumption that certain types of expressive phenomenological transformations of the world are always linked to the having of the thereby designated emotion. I would not though be able to come to the conclusion that I am sad *non-inferentially*, ie. on the sole basis of how the world appears to me. Although there may be a direct conceptual relation between the blue appearance of the world and sadness *in general*, there does *not* seem to be any direct conceptual link between the blue appearance of the world from my perspective and *my* being sad. [1]

Thus, to solve our puzzle about emotional self-knowledge, that is, the puzzle of how we are able to *self-ascribe* emotions directly on the basis of looking out at the world, this world will have to be shown to be transformed in emotion experience in a way implicit in which is already reference not just to a general type of emotion but specifically to ourselves as having an emotion of that type. Earlier on (in the first part of this chapter), it was pointed out that such things as pieces of music or colours or even forms of behaviour could be expressive of certain types of emotions in virtue of properties entirely *independent* of the mental states of any particular person, that is, in virtue of entirely 'objective' properties of these phenomena themselves – the type of music, the colour shade, the forms of behaviour, etc. If authoritative knowledge of our own emotions based directly on how the world seems to us is therefore to be possible, the phenomenological transformations of the world effected by (or in some other way associated with) our emotions must not just be emotionally expressive in the ways that pieces of music, lighting effects, or colours are. They must, that is, not just be directly expressive of specific types of emotions, but also directly (ie. not by way of any inference or identification assumption) expressive of *ourselves* as having these emotions.

Put differently, and as mentioned at the very beginning in the introduction to this chapter, the world will have to be shown to be not just *emotionally tinged* by our emotions, but emotionally tinged in implicitly self-world *relational* colours. The aim of the next chapter and final aim of this thesis will be to show this.

---

[1] A very similar problem arose, recall, for cognitivist theories of emotion. Apart from the general reasons there were for rejecting cognitivism as an account of what emotions *are*, a more specific problem arose also for cognitivism as providing a solution to our concerns about self-knowledge – ie. when taken merely as an account of how emotions find *expression* in consciousness. Much like in the present example of the world coming to be tinted in blue, it was noted that coming to experience the world, or particular aspects of it or particular objects in it, as having general emotion-specific properties such as of being 'frightening', 'desirable', 'lovable', 'praiseworthy', etc. was not enough to tell us, simply on the basis of looking out at the world, that *we ourselves* had the corresponding type of emotion, eg. fear, desire, love, etc.

Chapter 8: Emotions, desires and hodological space

The final task before us in this thesis is this: to arrive at a concrete account of how we are able to self-ascribe our own emotions directly on the basis of looking out at the world. Doing this however, we have seen, will require showing essentially two things – first, that and how having an emotion can be seen to transform or 'colour' the world as experienced from our perspective in a way that is emotionally *expressive*, and second, how it can be seen to do so in a way that is specifically expressive of *ourselves* as having an emotional attitude towards the world. Having done the first of these two things to some extent in the last chapter, our main challenge in this chapter will be to do the second.

To anticipate, this will be done in what follows essentially by appealing to the idea of our experiencing the world, in having emotions, in terms of 'hodological space' (ie. as making specific demands upon us, as *to-be-acted-upon* in specific ways) and/or as directly *acting upon us* in various ways – these being, we will see, (a) distinctively emotional and motivational ways of experiencing the world, and (b) ways that contain moreover implicit in them reference specifically to *ourselves* as having specific emotional or motivational attitudes towards it. Some insight into this, and support for the view that there are indeed such self-world relational elements in emotion experience, can I believe be found to start with in some of the recent psychological literature, and in particular in so called 'appraisal theories' of emotion. This chapter will therefore begin by first considering some of this literature. It will then move on to connect key claims made by appraisal theorists with discussions along similar lines in the philosophical, neuroscientific and other parts of the

psychological literature, aiming to ultimately show how these discussions can be appealed to to provide a concrete solution to our puzzle about the self-ascription of emotions.

8.1 The appraisal theory of emotions

In it's original form, the core suggestion of the appraisal theory of emotions pioneered by Lazarus and Arnold,[1] is that emotions (generally construed as episodic states or processes of some form – though not necessarily conscious ones) always arise somehow through the mediation of *appraisals* of external events (ie. as opposed to directly from the events themselves, or directly from physiological or neural causes), and in particular, through the mediation of appraisals of these events in relation to one's own concerns, priorities, desires, and felt coping potential. In other words, in the first instance, the appraisal theory of emotions is a theory of the *elicitation* and *differentiation* of emotional responses. It suggests that emotions are elicited by *appraisals* of events, and, that these eliciting appraisals (rather than the events appraised) are what determine what specific emotion as opposed to any other will ensue. Next, the appraisal theory makes also a very particular claim about the *kinds* of appraisals that are involved in this elicitation of our emotions, suggesting that they are essentially *relational* appraisals, ie. appraisals of events and circumstances in relation to our own interests, goals, and so on.

Now, the main advantage of, and motivation for, the adoption of this view over all other theories of the causes of emotions, was at its origin that by introducing the above kind of intermediate causal layer between external events and emotions, the view would be able to render intelligible, on the one hand, how the same events or external stimuli can sometimes elicit on different occasions (and across people and cultures) very different emotional responses (ie. in that different people may appraise the same events differently; they may have different concerns in relation to these events, etc.), and, on the other hand, how different events can elicit across people and populations sometimes the *same* emotional responses (different events may be of similar concern and importance to different people; they may therefore, although

[1] See for instance (Arnold 1960), (Lazarus 1982) and (Lazarus 2001 in Scherer, Schorr, Johnstone eds. 2001)

different, be evaluated similarly by them, etc) . How might this relate however to our concerns about self-knowledge?

At first sight, such a theory may seem to place the relational element in the entirely wrong place. That is, inasmuch as the appraisal theory is a theory essentially of the *causes* of our emotions and *not* therefore a theory of anything that is involved in actually *having* an emotion or in *experiencing* the world in having an emotion (which is what we are looking for), it is not immediately obvious how it might be of any direct relevance to our purposes, that is, how it might be able to provide us with an answer to the question of how looking out at the world in *having* an emotion can constitute direct evidence for how we now feel.

Not all versions of the appraisal approach to emotions however are theories strictly of the *causes* of emotions. In response in fact to a wide range of difficulties that have been raised (often by appraisal theorists themselves) against the appraisal theory in its original form, a number of different versions of the view (ones for that matter better supported by the available psychological data) have emerged, many of which in the end can, we will see, provide us with some valuable insights into how we are able to have first person authoritative access to our own emotions.[1] A number of the more recent versions of the appraisal theory actually maintain that appraisals are not just the causes of emotions but often also *constituents* of emotions and of emotion experiences. These versions of the theory may therefore have something to tell us about how the way the world is for us in *having* an emotion can contain implicit in it reference to ourselves as having it. In order to see how they might do so though, it is worth looking first at some of the problems with the theory in it's original form that these more promising versions have evolved to address.

A first difficulty faced by the appraisal theory in its original form (ie. in the form 'no antecedent appraisal – no emotion') is that, in a number of cases, emotions do not actually seem to be caused by antecedent appraisals. Many emotions seem to arise directly from external stimuli such as sudden positive or adverse circumstances. Anger or irritation for instance may on occasion arise directly from, say, someone's stepping on one's foot or from one's hitting one's head against the kitchen cabinet.[2] Fear may similarly arise directly from being suddenly attacked from behind; joy as a

[1] For a selection of different versions of the appraisal approach to emotions see (Scherer, Schorr, Johnstone eds. 2001)

[2] This example is taken from (Frijda & Zeelenberg 2001)

direct result of having taken some psychoactive drug; depression as a direct result of a naturally occurring chemical imbalance in one's brain, and so on. Of course, one could try to argue that even in these cases an intermediate appraisal *is* involved; only it is extremely quick and not conscious, and so for that reason appears not to exist. To insist though in this way on the existence of a mental process that is neither conscious nor in any other way manifest whether in one's behaviour or otherwise would it seems just be to assume without argument that the appraisal theory in it's strongest version is correct. To point this out is of course not to say that the emotions in the above examples may not sometimes arise *with* appraisals or *with* sudden changes in one's evaluations of the circumstances one is in. To insist however that the appraisals in question occur in these cases strictly *prior* to, and as the mediating *causes* of one's emotions seems somewhat unjustified.

Even taking for granted that at least in some cases particular appraisals (or particular patterns of appraisals) do occur prior to one's emotional responses and *are* causally responsible for the latter, a second problem arises. In such cases where eliciting appraisals are present, it is not clear that these appraisals always figure *solely* as the causes of one's emotions (ie. as the mere triggers of one's emotions, thus essentially ceasing at the point at which the emotion begins – or, put differently, ceasing at the point at which one begins to be disposed in the ways constitutive of having the relevant type of emotion). Most of the available psychological data suggests (as indeed does mere introspective reflection) that although certain appraisals may indeed occur *prior* to one's feeling a certain way and be causally *responsible* for how one feels, they may also, and often do, *persist* throughout the emotional response, constituting even an *essential* aspect of what it is to be feeling an emotion of that type – ie. being depressed for instance may be said to involve by its very essence appraising the world as 'uninviting', as 'devoid of interest', etc. Moreover, such emotion-*constituent* appraisals may it seems on occasion actually *differ* from the appraisals that initially triggered the emotion – eg. the appraisal that gave rise to one's depression may not be the same appraisal as that (or those) involved in one's now *being* depressed. Finally, other appraisals may be purely *consequents* of one's emotions, arising only subsequently out of one's current emotions, and perhaps constituting essentially part of *another* emotion – eg. one's feeling depressed may have led one to feel angry and to thereby appraise things in ways constitutive of being angry.

Thus, to sum up, not only may some emotions it seems be caused by things other than appraisals (eg. directly by external events or neural causes), but, even when appraisals are involved, they need not it seems always figure strictly as the *causes* of one's emotions. As pointed out by Frijda & Zeelenberg in fact, the actual empirical psychological data usually appealed to (ie. generally verbal reports) in support of the emotion-*antecedent* appraisal view does not actually allow for very clear disentangling between instances of antecedent, constituent, or consequent appraisals.[1] Appraisals may it seems be any, or all, of these three. That is, as put by Roseman & Smith, 'appraisals may be causes of emotions, components of emotions, as well as consequents of emotions',[2] a view now generally acknowledged by most appraisal theorists. Certain versions of the appraisal theory (ie. most of the more recent versions) may thus in the end actually contain some important insights for us into how the way the world seems to us in having an emotion might, as needed, provide us with a direct ground for self-ascribing this emotion – at least in that these versions of the appraisal theory are proposals not just about what *causes* our emotions, but also about what is involved in some cases in our actually *having* these emotions. Even so though, a third problem remains both with respect to our concerns about self-knowledge and for the appraisal approach more generally.

On the face of it, whether appraisals are claimed to be antecedents or constituents of emotions, the picture seemingly painted by appraisal theories (sometimes referred to as 'cognitive appraisal theories') can seem somewhat unrealistically cognitivistic. It seems to assume, that is, a far greater degree of complexity and sophistication to be involved in having emotions or in their elicitation, than actually *does* seem to be involved in most cases. For one thing, the terminology of 'appraisals' or of 'cognitive appraisal processes' immediately suggests that what is being claimed is that certain complex evaluative *judgements* or complex processes of *propositional reasoning* are involved in the elicitation of emotions or in the having of emotions. Yet, clearly, no such processes and no such fully articulated propositional evaluations are always involved, be it prior to having emotions or in the course of having these emotions. To insist otherwise would just seem to be to go against all available evidence and common sense, not to mention that, specifically in relation to

[1] See (Frijda & Zeelenberg 2001); see also (Frijda 1993) for a discussion of how some appraisals may even be post hoc *rationalisations* of how one felt or of what *made* one feel a certain way.

[2] (Roseman & Smith 2001, p. 15)

our concerns about self-knowledge, it would be to make very little progress on the already rejected cognitivist theory of the manifestation of emotions in consciousness discussed in chapter 6.

Sensibly however, very few appraisal theorists (Reisenzein being perhaps the only one)[1] actually have such a strictly cognitivistic understanding of appraisals in mind. Despite the misleading terminology of 'appraisals', most appraisal theorists use the notion of 'appraisal' in a far looser sense, meaning by it not an 'evaluative judgement' but only *some* form or other of 'stimulus coding'.[2] This coding may take any number of forms – propositional, analogue, conscious or unconscious. The crucial idea behind appraisal theories is not one regarding the level of cognitive sophistication that must be involved in the coding of the events perceived, but the idea that emotions involve (or are caused by) not just the perception of events or the reception of external stimuli, but that they essentially involve (or are caused by) a coding of these events, objects, or external stimuli in terms of, or in relation to, *one's own concerns, desires, aims, felt coping potential*, etc. This, moreover, is an idea held not solely in the psychological literature, but one shared by philosophers such as Sartre (and, as we will see, by neuroscientists and other psychologists), that life is not a matter of mere perception of mind-independent objects and events; we also have things to get done, aims, concerns, and feeling about various external objects and events around us, and we tend to see things directly in relation to these aims, desires, concerns, etc. – ie. as means or as obstacles to them, as open or closed paths, as possibilities or impossibilities, and so on.

As far as most appraisal theorists are concerned, the so called process of appraisal may be an entirely immediate and unconscious one,[3] or, where it is a *conscious* process, it need not be a fully articulated, propositional, one. Many appraisal theorists in fact go so far as to argue that when the process is conscious, the kinds of appraisals that are distinctive of having emotions and of their elicitation are essentially immediate, *non-propositional, perceptual* appraisals. Frijda for example stresses in a number of places,[4] that in actual fact, the kinds of appraisals that produce

---

[1] See (Reisenzein 2001)

[2] This term is taken from (Le Doux 1989)

[3] See (Frijda 2001); also (Roseman & Smith 2001)

[4] See for instance (Frijda 1988) and (Frijda & Zeelenberg 2001)

and sustain emotions or that are involved in actually *having* emotions or emotion experiences are specifically not 'appraisal *judgements*' but 'appraisal *experiences*'. He argues on the basis of experimental data that propositional information tends in fact to be emotionally *inert* in a way that analogue *experiential* content is not.[1] Evidence for this, he suggests, can be found for instance in experiments involving people suffering from phobic fears (ie. irrational or seemingly inappropriate or disproportionate fears). The data on the treatment of phobias, he points out, suggests that when it comes to curing sufferers of such fears, methods of live modelling (ie. methods involving various forms of direct perceptual confrontation with the objects of one's fear, or parallel confrontations in visual imagination, memory experience, etc. – trying to get people to *see* things differently) are far more effective than are more cognitive treatment methods (ie. those involving such things as trying to talk people out of their fears, say, by making them come to acknowledge that there is nothing of direct threat to them in small spaces, large crowds, spiders, feathers, etc). Most sufferers from phobias would *already* acknowledge without the need for talking into, that there is nothing dangerous or truly threatening about the objects of their phobic fears; it is rather that they cannot help *seeing* these things, somehow primitively (and in some *absolute* sense)[2] as just *to-be-avoided*, as *threatening one* (even if at a distance, without any obvious means), etc. As Frijda puts it, when it comes to emotion (though he admits to not being entirely clear about why this is so) 'knowing means less than seeing', 'words mean less than tone of voice', and, 'feeling means more than knowing'.[3]

This appears to be so even in less pathological cases. When someone accidentally steps on another's foot for example, this can immediately trigger a feeling of anger or hostility, manifesting itself (in part at least) in the perpetrator's striking one primitively as *to-be-blamed* or as *to-be-hit*, although one does not actually believe them to be blameworthy or deserving of any act of revenge. Similarly, in irritation, say, at someone blocking one's view at the cinema, one may come to experience this person primitively as *affronting one* although one does not

[1] See (Frijda 1988); also (Bridger & Mandel 1964); (Bandura 1977) and (Lang 1977) for the experiments themselves.

[2] By 'absolute' I mean not relative to some context, ie. not 'at time t' or 'in the present circumstances' or 'given that snakes are poisonous' etc.

[3] (Frijda 1988 p.275 in Jenkin, Oatley & Stein eds. 1998). See also (Frijda and Zeelenberg 2001)

believe them to be in any way even aware of one's presence behind them. In other words, the kinds of appraisals that seem to be most crucial to the having of emotion experiences and the making of which may even seem to entail that one *does* have the emotions in question, are essentially 'appraisal *experiences*' (ie. experiences of things directly *as* this or that, or *as* directly acting upon one, or *as* immediately calling for certain courses of action on one's part, etc.) rather than, and often independently from, 'appraisal *judgements*' (ie. judgements *that* things truly are one way or another, or that one *ought* to act in this way or that way, or that someone/something really is acting upon one in this or that way).

Along a similar vein, Magda Arnold also makes this point[1] (though she is motivated more specifically by the lack of control we seem to have over our appraisals in emotion) that appraisals in emotion (or in the elicitation of emotions) are essentially forms of perceptual *experiences*. Appraisals are, she stresses, emotionally 'intuitive' assessments of the here and now and not deliberative rational processes or reasoned judgements. Also along these lines, Lambie & Marcel make the point of clearly distinguishing between 'evaluative judgements' on the one hand and 'appraisal-*awareness*' of the world on the other, the latter of which they take to constitute a distinctive aspect of the way in which specifically *emotional* states manifest themselves in consciousness.[2] And indeed, intuitively, the link between appraisal awareness and emotions seems to be far tighter than that between appraisal judgements and emotions.

To put the same point by way of an example taken earlier from Sartre, although one could it seems *judge* that a face behind a window is threatening and that one ought to avoid it without being afraid, it is far less clear that one could directly *experience* the face primitively as terrifying, or directly as *threatening one*, or primitively as *to be avoided*, without actually feeling fear. The same point comes up also in an example by Goldie of how the world might appear to one upon looking down, in fear, from the edge of a steep cliff. He writes: 'The edge looms large in your gaze, and somehow seems to be pulling you towards it...' despite your not believing it to actually be pulling you anywhere nor being any larger that it is.[3] It is

[1] See (Arnold 1960)

[2] See (Lambie & Marcel 2000)

[3] (Goldie 2000 p. 76)

experiencing the world in this way that is crucial to the feeling of fear, he suggests, quite independently from what one believes.

Goldie's example, incidentally, is interesting in more ways than one. It illustrates first of all, as intended, this point about how the way one *experiences* the world can come apart from how one actually *believes* it to be, and how it is only the former which is truly essential to the feeling of an emotion (ie. one may believe *that* the ground below the cliff is pulling one towards it and that the edge is dangerous *without* feeling fear; but it does not seem that one could truly *experience* the edge as luring one towards it, or the ground as calling one to jump, or the open space primitively as to-be-leapt-into without actually feeling fear). This same example however also illustrates another point, made earlier on in this chapter, about how we often tend to describe our emotions in mixed world/self focused terms. The example reads more fully: 'The edge looms large in your gaze, and somehow seems to be pulling you towards it...you imagine yourself [...] suicidally throwing yourself over the cliff; you feel faint [...] you tremble, feel a damp sweat', etc. Goldie's manner of capturing the experience of fear is it seems a classic case of mixing somewhat indiscriminately descriptions of the way the *world* appears to one and descriptions of how one oneself *feels* in order to relate one's emotion, thereby suggesting again that certain ways the *world* appears to one can be as expressive (ie. as suggestive, evocative, etc.) of particular emotions as can be more explicit self-focused descriptions. Certain ways in which the world strikes us in emotion, that is, seem to be themselves already somehow distinctively *emotional*, or emotionally *expressive*.

To round up our discussion of appraisal theories, we have seen that many appraisal theorists, like many philosophers and other psychologists, when speaking of 'appraisals' or of 'appraisal processes' in emotion, actually have in mind appraisal *experiences* rather than appraisal *judgements*, and that they have these appraisal experiences in mind moreover not just as being the *causes* of emotions but often as being ways in which we experience the world from within our point of view in *having* emotions. Adding to this then the claim (shared by all appraisal theorists) that the appraisal experiences involved in (and distinctive to) emotion are essentially self/world *relational* appraisal experiences (ie. appraisals of the world in relation to our own *concerns*, *interests*, *desires*, etc. and in particular appraisals of the world as calling for certain actions from *us* and as acting upon *us* in various ways) emotional attitudes more generally), the appraisal approach to emotions can truly be seen to

point towards a picture of what is involved in having emotions very much along the lines of what we need to solve our puzzle about emotional self-knowledge. Recall, what we need is an account of the way in which (a) we *experience* the world in having an emotion, that is (b) implicitly self/world *relational*, and more specifically (c) implicit in which is a distinctively *emotional* self/world relation or *emotion*/world relation, not just a cognitive or perceptual one.

What remains now to be done is for us to get clearer about this idea of there being a way in which we experience the world that is directly relational upon our concerns, desires, felt coping potential, and emotional attitudes more generally. What might experiencing the world in this way actually amount to? Before that though, it is worth considering some further support (beyond that provided by appraisal theories) for this view that we *do* in fact experience the world in such an implicitly emotion/world relational way.

8.2 Solving the puzzle

That we experience the world in having emotions in a way that is distinctively relational upon our current concerns, aims and coping potential is held not just by appraisal theorists but comes up also in a number of other places, in the philosophical literature (in Sartre in particular), in various other parts of the psychological literature (in Lambie & Marcel for instance) but also strikingly in the neuroscientific literature, in particular in the literature regarding the role of our emotions in reasoning and in decision making.

Earlier on in this thesis,[1] it was mentioned that neuroscientists such as Damasio as well as other authors inspired by similar neuroscientific findings (eg. De Sousa) argue that what is important to us, what we are concerned about, what aims we have, and more generally how we feel about various things (whether we are afraid of x, in love with y, angry about z, etc.), determines to a great extent what our *attention* is turned to, or, as they put it (in more world-oriented terms), it determines the *salience* of certain things rather than others in our perceptual experiences, in our memory experiences, in our visual imagination, and so on. To the extent that this is so, it immediately points to already one clear way in which the phenomenology of

[1] See footnote on p.178 above

different people's experiences might differ in accordance with what they are concerned about, what they desire, and more generally with how they feel about various things. Different aspects of the same environment may, for instance, figure in the foreground of my attention and only in the background of your attention. Certain things that I am afraid of will be more salient from my perspective than from yours who are not afraid of these things and vice versa.

Of course, the mere fact that the phenomenological salience of certain things rather than others is determined by our emotions, desires, and concerns is not yet enough to show that the fact of our having these emotions actually spills into the conscious *content* of our experiences. But, as noted earlier in discussing the neuroscientific literature, this is not the only kind of phenomenological transformation pointed out by neuroscientists to be incurred by our emotions on the world from our perspective.

Both Damasio and De Sousa make the additional and in their view crucial point that depending on, say, whether we have a favourable or unfavourable attitude towards something, or depending on such things as whether we are afraid, angry, in love, guilt-ridden, etc. certain things will not just come to *stand out* phenomenologically more than others in our experience, but these salient aspects of the environment will come to stand out in specific *ways*, in particular, and amongst other things, as calling for certain courses of action on our part – eg. as immediately *to-be-avoided*, as *to-be-obtained*, as *to-be-kissed*, etc. Particular courses of action themselves may also come to strike us immediately as *the* ones to take, or as *not* to take and so on. In other words, on this view, the world comes to strike us directly, in having emotions, in ways that somehow already reflect various motivational attitudes constitutive of our emotions. Certain paths just immediately come to strike us as *to-be-followed*, certain others immediately as *to-be-avoided*, and so on, these all being ways of experiencing the world which, first of all, clash somewhat with our ordinary understanding of how the world really works (ie. on a physicalistic world picture, inanimate objects or paths do not make demands upon us) and, in so clashing, may come to be seized by us directly as expressive of ourselves as standing in some distinctively 'action-ready' relation to the world rather than of the world as literally pulling us into action.

Importantly, according to this same neuroscientific literature, and as was argued by appraisal theorists, the above ways in which we come to experience the

world in emotion are clearly *not* supposed to be ways in which we come to *judge* the world to be on the basis of reasons. Rather, they are supposed to be ways in which we *directly experience* the world, or in which the world directly *strikes* us. The crucial idea behind Damasio's and De Sousa's views is in fact that the essential function of our emotions in our mental lives, and in particular in our decision making processes, is that of *supplanting* processes of rational deliberation and *supplanting* reasoned judgements, thereby (a) saving us from getting caught up in inadaptively long deliberations in circumstances where coming to a quick decision is of the essence, and (b) saving us also from potential decision-making deadlocks in situations where pure reason would not favour either of, say, two equally acceptable options. When an unidentified object for instance comes quickly towards us, the attitude of fear and its constitutive manifestation in our experiencing the oncoming object primitively as *to-be-avoided*, is on this view supposed to save us from the unfortunate consequences that would follow from our stopping to reason for too long through the options. And, in situations of rational deadlock, a more favourable gut feeling towards, say, one of two routes equal in length to the bus stop (and its manifestation in our experiencing the one route as just *the* one to take) is supposed to prevent us from getting stuck in the absurd situation of deliberating forever about something so mundane.

If emotions are indeed to play the above roles in decision making (which they *do* seem to play in many cases – even if just as a matter of contingent empirical fact), they *must* result not just in our coming to make certain reasoned judgements in particular circumstances, but in our *directly experiencing* certain paths as *to-be-taken*, certain suspicious people as *to-be-avoided*, certain objects as *to-be-dodged*, and so on. This being so would moreover be very much along the lines of what we need to solve our puzzle about self-knowledge – ie. we need to identify ways in which the world *directly* strikes us in experience such that there might be implicit in it reference already to ourselves as standing in some distinctively emotional or motivational relation to it.

Damasio and De Sousa actually go somewhat further than just suggesting that emotions do, as a matter of contingent empirical fact, supplant our deliberative processes, ie. in the sense of just happening to determine our decisions *instead* of reason. They argue additionally that these decisions that our emotions lead us to make, and that manifest themselves as felt demands upon us from the world, are also the most *rational* and most *adaptive* decisions for us to make. In some cases, the idea

is that emotions actually lead us, somehow by their very nature, to the very *same* decisions that more lengthy deliberative processes of reasoning would do – only much faster and more effectively. And, in other cases, their idea is that without emotions manifesting themselves in our seeing certain things as just *to-be-done*, it would be *in principle impossible* for us to get out of particular decision-making deadlocks. Put differently, according to Damasio in particular, our emotions not only supplant reasoning in some cases but play (a) an *in principle* indispensable role in avoiding decision making deadlocks, and (b) constitute by their very nature a fast-track route to the most *rational* and most *adaptive* courses of action.

For a number of reasons however, which it is not entirely relevant to go into here, I do not believe that these strong claims can stand up to scrutiny. Very briefly, rational decision making deadlocks could it seems be avoided in a number of ways other than by the use of gut feelings or emotions. For example, one could toss a coin in some cases, apply an arbitrary rule in other cases (eg. 'always pick the item most to the left of you when choosing from identical copies of an object'), or apply a well tested *rationally* derived rule (eg. 'never speak to strangers in dark streets at night') in yet other cases. Furthermore, not all decision making deadlocks could it seems be avoided even *with* the use of emotions. Some decision making deadlocks may in fact be *caused* by emotions – eg. two simultaneous emotions may be pulling one in opposite directions – and require *reason* this time to be broken out of. Finally, it is also far from clear that doing what 'feels right' will always lead to the most rational (only faster) or most adaptive courses of action. Phobic fears, depression, overexcitement, etc. may all lead one, quite on the contrary, to highly *ir*rational, *in*adaptive, or disproportionate courses of action.[1]

Despite these shortcomings of De Sousa's and particularly Damasio's views, what seems difficult to deny is that, as matter of contingent empirical fact, our emotions *do* to some extent make us see the world in the ways described by them, regardless of how rational or adaptive this might be in particular cases. Our emotions do make certain paths just come to be seen by us as *to-be-followed*, others as *to-be-avoided*, shoes as *to-be-resoled*, a partner primitively as *humiliating us* or an

[1] It may of course be that many emotions are evolutionary adaptations, and so states which led to adaptive behaviour in the environment in which they evolved. This does not however mean that having these emotions in the present environment (or for that matter in all instances in the past environment) will always lead one (let alone do so *in principle*) to act in the most rational or adaptive ways.

adversary as *to-be-punished*,[1] something we long for as *to-be-obtained*, someone we are angry at as *to-be-blamed* or as *to-be-hit*, someone we are attracted to as *to-be-touched* or as *to-be-kissed*, etc. If our emotions did not make us experience their objects in these ways, much of our behaviour (eg. in cases where we did not in fact toss a coin, or have a ready rule to apply, or have the time to engage in rational deliberation) would be difficult to account for. And, this is a view supported to varying degrees of explicitness not just by neuroscientists, but as we have seen by appraisal theorists and a number of other psychologists and philosophers as well.

Both Lambie & Marcel and Sartre, to introduce some of their technical jargon, describe the above type of phenomenology of emotion experience as experiencing the world in terms of a form of 'action' space, or as they put it '*hodological* space' whereby the world is not experienced just as a world of objects and events distinct from ourselves bearing various general properties, but also as a usable world, one affording various possibilities for action,[2] and crucially as a world not just *affording* possibilities for action but *requiring* certain things of us, making specific *demands* upon us, having various *imperatives* for us attached to it (as a world of paths experienced as *open* or *closed-off*, or indeed as *to-be-followed* or as *to-be-avoided*, etc.), these all being reflections of various motivational and emotional attitudes we might have towards the world. A situation may be felt not just as *avoidable* but as *to-be-avoided*; a person not just as a *possible target* but as *to-be-attacked*; a course of action (even an unreasonable one, eg. washing one's hands continuously) as just *to-be-carried-out*, and so on. This way of experiencing the world distinctive to emotion, to introduce another piece of technical jargon, is also sometimes referred to as 'gerundival'. A gerundival experience is, put simply by Lambie and Marcel, 'an experience a subject has of an object whereby the object strongly implies or *impels* an action that should be performed with regard to itself'.[3] Examples of such ways of perceiving the world are essentially cases such as those already mentioned of directly experiencing, say, a pet as 'to be stroked', a cake as 'to be eaten', a person as 'to be kissed' or as 'to be attacked', shoes as 'to be re-soled', and various other things as simply *to-be-acted-upon* in various specific ways.

---

[1] For a detailed discussion of the example of jealousy (ie. an example of a complex, multifaceted emotion) see (Goldie 2000, ch.8)

[2] This is suggested for instance by the theory of 'ecological optics' (Gibson 1979)

[3] (Lambie & Marcel 2000, pp.66)

Crucially, all the authors who discuss this (amongst which those mentioned here, ie. Sartre, Lambie & Marcel, Damasio, De Sousa, and many appraisal theorists) take these distinctively emotional or motivational ways of seeing the world, ie. in terms of 'hodological space' or 'gerundival perception', not just to be cases of experiencing aspects of the world as having *general* gerundival properties or as having *general* or *in principle* imperatives attached to them, but as demanding things specifically of *ourselves*. That these requirements experienced as imposed on us by the world contain an implicit or parenthetical 'by me' is stressed for example quite explicitly by Sartre, as well as by Lambie & Marcel, and is also clearly manifest in some of the central claims made by Damasio and De Sousa, particularly in relation again to the role that experiencing the world in this way is supposed to play in decision making.

For example, if as suggested by Damasio and philosophers such as De Sousa emotions are to play an effective role in supplanting lengthy deliberative processes, and in getting us immediately to act in effective ways in situations of rational deadlock or in cases where time is of the essence, it *must* be that our coming to experience, say, an oncoming object as to be avoided, actually be a case of our directly experiencing the oncoming object not just as to be avoided *in general*, but as to be avoided in a particular *way*, ie. by our *moving*, rather than, say, by our moving someone else. Similarly, in cases of desires that manifest themselves in our experiencing someone as, say, *to-be-kissed* or as *to-be-hit*, etc. the demand *must* contain an implicit 'by *me*' if it is to immediately give rise (without the need for reasoning or further premises) to the appropriate action. The desires or demands in question would not be satisfied if not effected by oneself. And, incidentally, it never seems to happen that one is left wondering whether a demand experienced as imposed by the world in emotion experience is actually made upon oneself or on someone else. The fact that the demand is made upon oneself is implicit already in the type of action required of one, ie. that of *moving* or *kissing* rather than that of moving someone else or getting someone else to kiss one's beloved.

In this way of experiencing the world (ie. as making certain direct demands upon *us*, as having immediate imperatives for *us* attached to it) is clearly a way in which we might directly experience the world from our point of view in having emotions that is (a) distinctive to emotion experience, and (b) distinctively expressive of *ourselves* as standing in a specifically motivational relation to the world. In

essence, the idea that can be derived from the above literature here is that, in emotion, the world comes to be transformed not just into a world of objects bearing various *general* gerundival properties – ie. with certain things being experienced as *to-be-avoided* generally speaking, particular people as *to-be-hit* on general principle, particular objects as *to-be-obtained* in the sense of being 'worth obtaining' – but rather, into a world of objects having imperatives distinctively for *oneself* attached to them. This is precisely what we need to solve our puzzle about authoritative self-ascription of emotion.

Furthermore, moving on now somewhat beyond emotion-induced experiences of the world in terms of 'hodological space', another closely related (and similarly relevant to our purposes) way in which the world can be seen to be phenomenologically transformed by our emotions, highlighted by Sartre and Lambie & Marcel as well as by other authors in the philosophical and psychological literature discussed here (eg. Goldie and Frijda amongst others), is by coming to be experienced not just as *demanding* actions of us, but also in turn as directly *acting upon us*, and in particular as acting upon us in what Sartre would call 'magical ways'[1], that is, often in entirely *unmediated, absolute* ways. Inanimate objects may for instance come to be experienced as *luring us* towards them or as *repelling us* at a distance. Mere 'situations' may come to be experienced as physically *pressuring us* or *suffocating us*, and so on. This need not be taken as a pathological way of seeing the world (except in extreme cases, where one is so caught up in one's emotional experiences that one comes to *believe* them), but rather, as Sartre suggests, as just one of the many ways in which we apprehend the world, this time a way of apprehending the world distinctive to emotion rather than mere cognition, and again a way of experiencing the world which contains implicit in it reference to ourselves as standing in a particular relation to the world (or it to us).

As mentioned earlier, and very much in line with Sartre's position but also with the many other views explored in this chapter, we do not it seems just perceive things or receive information from the world; we do not just experience the world or 'code it' in a detached physicalistic way (ie. as a world of objects bearing various general properties and interacting roughly in accordance with the laws of physics).

[1] By 'magical' ways we can understand ways which clash with our ordinary physicalistic understanding of how the world works and of what is possible.

We also have things to get done, particular concerns, interests, aims, feelings, and, as a consequence, we tend to see things (at the conscious level or at a more unconscious level, explicitly or implicitly) directly as means or obstacles to these aims and interests, as possible or impossible options, as open or closed paths, as inviting/ uninviting/ alienating situations, as situations to-be-avoided or to-be-leapt-into, that is, as having various other specific imperatives for us attached to them, and indeed as *acting* upon us directly in various ways – as pressuring us, as affronting us, as luring us, and so on.[1] To put things differently, we seem to experience the world, or specific events, circumstances, etc. not just in a detached way but often very much in relation to *ourselves*, as *addressing* themselves to us, as *requiring* things of us, as *acting* upon us, and generally as very much revolving around ourselves. This, it seems, is the specific contribution that our emotions make to our experience of the world. These ways of experiencing the world can moreover be seen to be in a sense *directly expressive* of ourselves as having specific emotional attitudes towards it, thus also beginning to make sense, as needed for our purposes in this thesis, of how we might be able to self-ascribe our own current emotions directly on the basis of looking out at the world, ie. directly on the basis of the way the *world* appears to us phenomenologically.

One question still remains though, namely that of why we tend to read the contents of our emotional experiences as directly expressive of ourselves as having particular emotional attitudes towards the world *rather* than as expressive of the world as actually having strange powers over us. What makes such ways of seeing the world directly suggestive, from our perspective, of ourselves as having particular emotions, rather than suggestive of the world as actually revolving around ourselves, that is, as actually being the way it seems in emotion experience, ie. the face at the window as

[1] I have been appealing to a great extent to Sartre's intuitions about the phenomenology of emotional experiences here, and there are indeed some obvious affinities between the present proposal and Sartre's position in (Sartre 1962). As has probably become clear by now though, these affinities hold more in spirit than in substance. Amongst Sartre's views which are not, and *need* not, be endorsed on the present picture are for example (a) his view that experiencing the world as, say, terrifying or as acting upon us in some way involves or implies actually *believing* it to be so; (b) his view that apprehending the world in these ways is what having an emotion *is*; or (c) his view that emotional 'transformations' of the world into the 'magical' are essentially *intentional strategies* to cope with a 'difficult' world. The latter view would in fact go directly against the present understanding of these phenomenological transformations of the world as cases of experiencing it in ways that are *non-intentionally expressive* of an emotion – not to mention that, taking these transformations as strategies to cope with a difficult world would also give rise to a number of other problems not faced by the present proposal, such as that of how to make room for cases of *positive* emotional transformations of the world, in particular when the world is not *difficult*. Having said that, Sartre does provide some insightful descriptions of the phenomenology of emotion experience, which, taken purely as such, *can* be appealed to and endorsed without having to take on any of his more controversial claims.

actually acting upon us at a distance, the person we love as actually having cast a spell on us, the ground below the cliff as actually calling us to jump etc? How is it that we do not, that is, like in cases of perceptual hallucination for instance (where the world is also somehow phenomenologically transformed), tend to take our emotionally transformed experiences at face value?

Well, in some cases at least, we *do* it seems take these experiences at face value. Error is clearly not ruled out, nor is it intended to be, on the present account of how we come to know our own emotions. The frequency of error, recall, was one of the crucial differences between our knowledge of our own emotions and our knowledge of our own cognitive states that we needed to make room for in solving our puzzle about mental self-ascription for the case of emotions. And, on the account put forward here, error is indeed possible.

One can on occasion get so caught up in an emotion that the way the world is experienced from one's emotional point of view is indeed confused with the way the world actually *is*. In what we would generally consider to be pathological cases, we might get so caught up in, say, jealousy, fear, anger, lust, etc. that we might loose touch with our more detached physicalistic understanding of reality and come to *believe* that things actually are the way our emotions present them to us – eg. a person we are in love with as having cast a spell on us, a partner we are jealous of as actually betraying us although we have no evidence for it, and in paranoid fear, against all evidence again, a person as actually out to get us and so on. These are all highly pathological cases though. Usually, the degree of projection is not so extreme. It may however still occur to some lesser extent in ordinary cases, and so lead one to be accordingly more or less inclined to either attribute emotions to oneself (say, anger or anxiety), or to self-ascribe primarily physical symptoms alongside making evaluative judgements about the world.

Moving away from pathological cases though, failures to correctly identify what emotions we have may also arise, on the present account, from a number of other factors, such as from the fact of the existence of *mixed* emotions, whereby the world comes to strike us in many different ways at once, in ways that may not necessarily fall under a single emotion category, or as making conflicting demands upon us, ie. in ways expressive of different emotions towards the same object simultaneously, and so on. This may make it in the end quite difficult in many cases

for us to judge immediately and with accuracy, simply on the basis of attending to the objects of our emotions, that we have particular emotions rather than others.

Finally, in cases of emotions that are highly structured and complex, eg. guilt about feeling angry about y, surprise at not feeling happy about something we expected to feel happy about, etc. the process of uncovering our emotions may require much more time and effort than the process of seizing certain other patterns (eg. very simple patterns of transformations of the world, or ones that we have experienced many times before) as directly expressive of specific emotions. That is, there is the phenomenon to be taken into account of emotions manifesting themselves not singly, but in a *wide range* of states, and often in particularly complex, multifaceted and structured ways. Grasping a particular complex transformation of the world (eg. where certain aspects strike one as to be acted upon in various ways, certain other aspects as acting upon *us* in various ways, etc.) as a manifestation of a single type of emotion might thus often be far from immediate, tending to be perhaps most immediate in cases where the same patterns have been experienced many times before, thereby eventually coming to be seized by one immediately as reflecting a single unified emotion.

As mentioned above however, the fact that so many forms of error are possible on the present picture of how we come to know our own emotions is no disadvantage of this view. It is on the contrary a *virtue* of the present proposal that it is able to accommodate the high susceptibility to error that is characteristic of our knowledge of our own emotions, and an even greater virtue of it that it can accommodate this and explain it *without* having to reject the view that we base our judgements regarding how we feel in a wide range of cases directly on the way the *world* seems to us. It even turns out, on this account, that it is precisely *because* of the sometimes complex ways in which our emotions manifest themselves in the way the world seems to us, that the process of self-ascribing our own emotions can often be a lengthy one and one particularly prone to error. At the same time, it is also on this account precisely because we self-ascribe many of our own emotions on the basis attending to the world, that our self-ascriptions of emotions can often be far more immediate and authoritative than our ascriptions of emotions to others. Self-ascribing our own emotions on the basis of how the *world* appears to us remains something that others, not directly seeing the world from our perspective, are not in a position to do. Our knowledge of our own emotions thus remains accounted for as being first-person

authoritative (or at least different in kind from our knowledge of the emotions of others) in a wide range cases, and indeed as being so much in the same way as is our knowledge of our own cognitive states, yet *without* leaving unaccounted for the important differences that do clearly exist between our knowledge of our own cognitive states and our knowledge of our own emotions.

To tie together the central claims made here about the way in which we experience the world in emotion, what has come to light is that our emotions can, following recent discussions in the psychological, neuroscientific and philosophical literature, be seen to manifest themselves in our conscious world-directed experiences (a) in a way that involves implicit or pre-reflective reference directly to ourselves as standing in a particular *relation* to the world (or the world to us); (b) in a way that is distinctive to *emotion* experience (only in emotion does the world sometimes take on a 'magical dimension', ie. appearing as acting upon us in various ways, as making specific *demands* upon us, etc); and (c) recognizably as being a way of experiencing the world that originates to a great extent from *ourselves* and from our own *attitudes* towards it (in that it is often in great *conflict* with our beliefs about how the world really works – ie. about it being a physicalistic world of objects bearing various general properties and not truly revolving around ourselves, not truly making demands upon us, not truly acting upon us directly in ways we don't take to be physicalistically possible, and so on); and, in the end (d), in a way that (possibly as a developmental result of all of the above factors) actually strikes us *directly* in a way that is immediately *expressive* or *suggestive* or *evocative*, from our point of view, of ourselves as feeling certain specific emotions towards it.

Put succinctly, this chapter has shown that and how a concrete account *can* be put forward of how self-ascribing our own emotions on the basis of attending to the objects of our emotions out in the world is possible, thereby ultimately also enabling the traditional problem of self-knowledge for emotions, and not just for cognitive states, to be solved, ie. the problem of accounting for the special access we seem to have to many (even if not all) of our own emotions. More than that, in appealing to key aspects of the recent psychological, neuroscientific and philosophical literature on emotion and emotion experience, this chapter has not only allowed us to make more concrete sense of our earlier conclusion that world-directed consciousness must involve an implicit form of self-consciousness, but also shown this conclusion to converge with, and be very much backed up by, discussions and data entirely

independent from the theoretical considerations about self-knowledge initially driving it in this thesis.

To conclude, this investigation of self-knowledge has brought to light essentially two things. First, on a purely theoretical level, it has shown in the initial chapters that the special kind of self-knowledge we have *must*, given its distinctive features, be rationally based on the way the *world* or the *objects* of our self-ascribed mental states appear to us from within our own conscious outward-looking point of view. Then, more concretely in subsequent chapters, it has suggested how evidence of ourselves as having the relevant mental states can indeed be seen to be present already implicitly in the way the world strikes us at the first-order level – both in cognition and in emotion. Drawing these two strands together, that is, on the one hand our theoretical conclusions regarding the possibility and nature of introspective self-knowledge, and on the other our suggestions regarding the phenomenology of world-directed consciousness, this thesis has shown that, and how, the path towards understanding immediate authoritative self-knowledge lies ultimately in a deeper understanding of world-directed consciousness, and in particular of world-directed consciousness as involving at the same time an implicit form of self-consciousness.

Bibliography

Armstrong, D.M. (1968), *A Materialist Theory of Mind.* London: Routledge and Kegan Paul

Arnold, M. B. (1960), *Emotion and Personality: Vol 1 Psychological Aspects.* New York: Columbia University Press

Austin, J.L. (1962), *Sense and Sensibilia.* Oxford: Oxford University Press

Baldwin,T. R. (1998), 'Objectivity, Causality and Agency', in Bermudez, Marcel and Eilan (eds.) *The Body and the Self.* Cambridge, MA: MIT Press

Bandura, A (1977), *Social Learning Theory.* Englewood Cliffs, NJ: Prentice Hall

Bermudez, Marcel and Eilan (eds.) (1998), *The Body and the Self.* Cambridge, MA: MIT Press

Bermudez, Jose-Luis (1998), *The Paradox of Self-Consciousness.* Cambridge, MA: MIT Press

Bilgrami, Akeel (1998), 'Self-Knowledge and Resentment', in Wright, Smith and MacDonald (eds.) *Knowing Our Own Minds.* Oxford: Clarendon Press

Block, Ned (1995), 'On a Confusion About a Function of Consciousness', in *Brain and Behavioural Sciences* (1995)

Boghossian, Paul (1998), 'Content and Self-Knowledge', in *Philosophical Topics 1989*

Bonjour, L. (1979), 'Can Empirical Knowledge Have a Foundation?', in *American Philosophical Quarterly* 15 (1978) 1-13

Bonjour, L. (1985), *The Structure of Empirical Knowledge.* Cambridge, MA: Harvard University Press

Bridger, W. H. and Mandel, J. J (1964), 'A Comparison of GSR Fear Responses Produced by Threat and Electrical Shock', in *Journal of Psychiatric Research,* 2, 31-40

Budd, Malcolm (1985), *Music and the Emotions: The Philosophical Theories.* London: Routledge and Kegan Paul

Burge, Tyler (1988), 'Individualism and Self-Knowledge', in the *Journal of Philosophy 1988*

Burge, Tyler (1996), 'Our Entitlement to Self-Knowledge', in *Aristotelian Society* Supplementary Volume 1996

Burge, Tyler (1998), 'Reason and the First Person', in Wright, Smith and MacDonald (eds.) *Knowing Our Own Minds.* Oxford: Clarendon Press

Campbell, John (1994), *Past, Space and Self.* Cambridge, MA: MIT Press

Cannon, W. B. (1927), 'The James-Lang Theory of Emotion', in *American Journal of Psychology* 39 (1927) 106-124

Cannon, W. B. (1929), *Bodily Changes in Pain, Hunger, Fear and Rage.* New York: Appleton

Churchland, P.M. (1991), 'Eliminative Materialism and the Propositional Attitudes', in D. Rosenthal (ed) *The Nature of Mind.* Oxford: Clarendon Press

Crane, Tim (1992), 'The Non-Conceptual Content of Experience' in Crane (ed.) *The Contents of Experience.* Cambridge: Cambridge University Press

Crane, Tim (1995), *The Mechanical Mind.* London: Penguin

Damasio, Antonio (1994), *Descartes' Error: Emotion, Reason and the Human Brain.* New York: G.P. Putnam

Davidson, Donald (1987), 'Knowing One's Own Mind', in *The Proceedings and Addresses of the American Philosophical Association,* 60 (1987) 441-58

Dennett, D (1978), 'Two Approaches to Mental Images', in his *Brainstorms*. London: Penguin

Descartes, Rene (1912), *A Discourse on Method, Meditations and Principles*. Toronto: Dent

De Sousa, Ronald (1987), *The Rationality of Emotion*. Cambridge, MA: MIT Press

Eilan, Naomi (1997), 'Objectivity and the Perspective of Consciousness', in *European Journal of Philosophy* (1997)

Evans, Gareth (1982), *The Varieties of Reference*. Oxford: Clarendon Press

Evans, Gareth (1985), 'Things Without the Mind', in his *Collected Papers*. Oxford: Clarendon Press

Fricker, Elizabeth (1998), 'Self-Knowledge: Special Access versus Artefact of Grammar – A Dichotomy Rejected', in Wright, Smith and MacDonald (eds.) *Knowing Our Own Minds*. Oxford: Clarendon Press

Frijda, N. H. (1988), 'The Laws of Emotion', in *American Psychologist*, 43 (1988) 349-358 reprinted in Jenkins, Oatley and Stein (eds.) (1998) *Human Emotions: A Reader*. Oxford: Blackwell Publishers, pp.270-287 (page references to latter)

Frijda, N. H. (1993),'The Place of Appraisal in Emotion', in *Cognition and Emotion* (1993) 357-387

Frijda, N. H., Kuipers, P. and ter Schure (1989), 'Relations Among Emotion, Appraisal, and Emotional Action Readiness', in *Journal of Personality and Social Psychology*, 57, 212-228

Frijda, N. H. and Zeelenberg, Marcel (2001), 'Appraisal: What is the Dependent?' in Scherer, Schorr and Johnstone (eds.) (2001) *Appraisal Processes in Emotion*. New York: Oxford University Press, pp.141-156

Gibson, E. J. (1969), *Principles of Perceptual Learning and Development*. New York: Appleton-Century-Crofts

Gibson, James J. (1979), *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin

Goldie, Peter (2000), *The Emotions: A Philosophical Exploration*. Oxford: Clarendon Press

Goldman, Alvin (1978), 'What is Justified Belief' in Pappas and Swain (ed.) *Essays on Knowledge and Justification*. Ithaca: Cornell University Press

Gordon, R (1980), 'Fear', in *Philosophical Review 89* (1980) pp. 560-78

Greenspan, P. (1988), *Emotion and Reasons: an Inquiry into Emotional Justification*. London: Routledge and Kegan Paul

Grice, H. P. (1957), 'Meaning', in *Philosophical Review 66 (1957) 377-388*

Grice, H. P. (1989), *Studies in the Way of Words*. Cambridge: Harvard University Press

Heal, Jane (1994), 'Moore's Paradox: A Wittgensteinian Approach', in *Mind* (January 1994)

Humberstone, I. L. (1992), 'Direction of Fit', in *Mind* 101 (1992) 52-83

Hume, David (1988), *Treatise of Human Nature*. L.A. Selby-Bigge (ed), Oxford: Clarendon Press.

Ishiguro, H. (1966), 'Imagination', in Williams and Montefiore (eds.) (1966) *British Analytical Philosophy*. London: Routledge and Kegan Paul

James, William (1884), 'What is an Emotion', in *Mind* 9 (1884)188-205

Jenkins, Jennifer M; Oatley, Keith and Stein, Nancy L. (eds.) (1998), *Human Emotions: A Reader*. Oxford: Blackwell Publishers

Kant, Immanuel (1929), *Critique of Pure Reason*. London: Macmillan Press

Kosslyn, S. M. (1995), 'Mental Imagery', in Kosslyn and Osherson (eds.) *An Invitation to Cognitive Science: Visual Cognition, Volume 2,* second edition, Cambridge, MA: MIT Press

Lambie, John and Marcel, Anthony (2000), 'The Varieties of Emotions Experience', *unpublished*

Lambie, John and Marcel, Anthony (2002), 'Consciousness and the Varieties of Emotion Experience: A Theoretical Framework' in *Psychological Review* 109 (2002) 219-259. US: American Psychological Association

Lang, P. (1977), 'Imagery and Therapy: an Information Processing Analysis of Fear', in *Behaviour Therapy*, 8, 826-886

Lazarus, Richard (1982), 'Thoughts on the Relations Between Emotion and Cognition', in *American Psychologist* 37 (1982) 1019-1024

Lazarus, Richard (2001), 'Relational Meaning and Discrete Emotions', in Scherer, Schorr and Johnstone (eds.) *Appraisal Processes in Emotion.* New York: Oxford University Press, pp. 37-67

Le Doux, J. E. (1989), 'Cognitive-Emotional Interaction in the Brain' in *Cognition and Emotion* 3 (1989) 267-289

LePoidevin and MacBeath (ed.) (1993), *The Philosophy of Time.* Oxford: Oxford University Press

Lormand, Eric (1996), 'Nonphenomenal Consciousness', in *Nous* (1996)

Martin, M.G.F. (1998), 'An Eye Directed Outward', in Wright, Smith and MacDonald (eds.) *Knowing Our Own Minds.* Oxford: Clarendon Press

Martin, M.G.F. (1999), 'Desire in Time', *unpublished*

McDowell, John (1994), *Mind and World.* Cambridge, Massachusetts: Harvard University Press

McDowell, John (1998), 'Response to Crispin Wright', in Wright, Smith and MacDonald (eds.) *Knowing Our Own Minds.* Oxford: Clarendon Press

Mellor, D. H. (1981), *Real Time.* Oxford: Oxford University Press

McTaggart, J. M. E. (1927), 'The Unreality of Time', reprinted in LePoidevin and MacBeath (eds.) (1993), *The Philosophy of Time.* Oxford: Oxford University Press

Moore, G. E. (1942), 'Reply to My Critics', in P. Schilpp (ed.) *The Philosophy of G. E. Moore.* LaSalle, Ill: Open Court

Moore, G. E. (1944), 'Russell's Theory of Description' in p. Schilpp (ed.) *The Philosophy of Bertrand Russell.* LaSalle, Ill: Open Court

Nagel, Thomas (1991), 'What Is it Like to Be a Bat?', in D. Rosenthal (ed) *The Nature of Mind.* Oxford: Oxford University Press

O'Shaughnessy, Brian (1980), *The Will: A Dual Aspect Theory.* Cambridge: Cambridge University Press

Peacocke, Christopher (1992), *Study of Concepts.* Cambridge, Massachusetts: The MIT Press

Peacocke, Christopher (1996), 'Entitlement, Self-Knowledge and Conceptual Redeployment', in *Aristotelian Society* Supplementary Volume (1996)

Peacocke, Christopher (1998), 'Conscious Attitudes, Attention and Self-Knowledge', in Wright, Smith and MacDonald *Knowing Our Own Minds.* Oxford: Clarendon Press

Plantinga, Alvin (1993), *Warrant: The Current Debate.* Oxford: Oxford University Press

Prior, Arthur (1962), 'Changes in Events and Changes in Things', reprinted in LePoidevin and MacBeath (eds.) (1993), *The Philosophy of Time.* Oxford: Oxford University Press

Reisenzein, Rainer (2001), 'Appraisal Processes Conceptualized from a Schema-Theoretic Perspective: Contributions to a Process Analysis of Emotions' in Scherer, Schorr and Johnstone (eds.) *Appraisal Processes in Emotion.* New York: Oxford University Press

Roseman, Ira J. and Smith, Craig A (2001), 'Appraisal Theory: Overview, Assumptions, Varieties, Controversies', in Scherer, Schorr, and Johnstone (eds.) (2001) *Appraisal Processes in Emotions.* New York: Oxford University Press, pp. 3-19

Rosenthal, David (1991), 'Two Concepts of Consciousness', in Rosenthal (ed.) *The Nature of Mind.* Oxford: Oxford University Press

Russell, James (1998), 'At Two With Nature: Agency and the Development of Self-World Dualism', in Bermudez, Marcel and Eilan (eds.) *The Body and the Self.* Cambridge, MA: MIT Press

Ryle, Gilbert (1966), *The Concept of Mind.* Harmondsworth: Penguin

Sartre, Jean-Paul (1943), *L'Etre et le Neant.* Editions Gallimard

Sartre, Jean-Paul (1969), *Being and Nothingness.* London: Routledge

Sartre, Jean-Paul (1971), *Esquisse d'une Theorie des Emotions.* Paris: Hermann

Sartre, Jean-Paul (1962), *Sketch for a Theory of the Emotions.* London: Methuen

Scherer, Klaus R; Schorr, Angela and Johnstone, Tom (eds.) (2001), *Appraisal Processes in Emotion.* New York: Oxford University Press

Shoemaker, Sydney (1988), 'On Knowing One's Own Mind', in *Philosophical Perspectives, 2, Epistemology* (1988)

Shoemaker, Sydney (1996), 'Self-Knowledge and 'Inner-Sense'', in Shoemaker, S. *The First Person Perspective and Other Essays.* Cambridge: Cambridge University Press

Shoemaker, Sydney (1996),'Introspection and the Self', in French, Vehling and Wettstein (eds.) *Studies in the Philosophy of Mind.* (Midwest Studies in Philosophy, Minneapolis 1996)

Smith, Michael (1987), 'The Humean Theory of Motivation', in *Mind* (1987) 96 31-61

Sobel, David and Copp, David (2001), 'Against Direction of Fit Accounts of Belief and Desire', in *Analysis* 61 (1) (2001) 44-53

Solomon (2003), *Not Passion's Slave.* New York: Oxford University Press

Spinoza, Baruch (1985), *Ethics*, in E. Curley (ed.) *Collected Works of Spinoza*, 2 volumes. Princeton

Strawson, P. F. (1959), *Individuals.* London: Routledge, chapter 2

Strawson, P.F. (1966), *The Bounds of Sense.* London: Routledge

Tye, Michael (1991), *The Imagery Debate.* Cambridge, MA: MIT Press

Williams, Bernard (1978), *Descartes: The Project of Pure enquiry.* London: Penguin

Wittgenstein, Ludwig (1953), *Philosophical Investigations.* Oxford: Blackwell

Wollheim, Richard (1999), *On the Emotions*, New Haven: Yale University Press

Wright, Crispin (1998), 'Self-Knowledge: The Wittgensteinian Legacy', in Wight, Smith and MacDonald (eds.) *Knowing Our Own Minds.* Oxford: Clarendon Press